ISSN 0494-7304 0132-053x

ARTU ÜLIKOOLI

TOIMETISED

ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

968

STOCHASTIC PROCEDURES

TARTU 1994

TARTU ÜLIKOOLI TOIMETISED
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
Alustatud 1893. a. VIHIK 968

STOCHASTIC PROCEDURES

Tartu 1994

Vastutav toimetaja Tõnu Kollo

TÜ toimetised
Vihik 968
STOCHASTIC PROCEDURES
Tartu Ülikool
EE 2400 Tartu, Ülikooli 18
T. 83. 330. 6,5.
TÜ trükikoda. EE2400 Tartu Tiigi 78

Contents

Acta et Commentationes Universitatis Tartuensis, 968 ,1994, pp. 5-16.

An Example of Uniform Strong Laws of Random Lines

Tarmo Koll

Summary. A result is proved for random lines by which the mean length of a random chord of a planar random figure converges uniformly (over the set of figures) to its expectation with probability one. Then an example of an uniformly and strongly consistent estimator of the area of a planar figure is given.

Key words: *integral geometry, random lines, empirical processes.*

Introduction

Integral geometry provides us with a great variety of formulas which show the linkage between different parameters of geometrical objects. In many cases some of these parameters, like the areas of figures and the lengths of curves, are associated with the expectations of certain random variables. Estimation of these expectations gives us also a possibility to estimate the related geometrical parameters (see, for example, Santalo (1976), pp. 31, 218).

Our main concern in this paper is with a large class of figures on the plane and we focus on finding an estimator of the area of these figures which is uniformly (over the class) consistent with probability 1.

The essential part of this paper is section 2, where we first prove a uniform strong convergence theorem for random lines, and then derive a uniformly and strongly consistent (USC) estimator to the area of a planar figure with the help of random line segments.

1. Some facts from integral geometry

First we will give a short glance at the formulas of integral geometry which we need. All the basic formulas can be found in

any book on this subject (Santalo (1976), for example).

In integral geometry a line g on the plane is usually given by parametrization on $[0,\infty) \times [0,2\pi)$ via the identification

$$g = g(p,\varphi) = \{(x,y) : x \cdot \cos\varphi + y \cdot \sin\varphi - p = 0\}. \qquad (1)$$

The measure of a set G of lines g is defined as integral over the density of lines $dG = dpd\varphi$. Up to a constant factor this density is the only one which remains invariant under the group of translations and rotations in R^2. So we will consider the lines g which are uniformly distributed over the whole plane.

If we restrict our attention only to the lines which intersect a convex set K, we have that the measure of such set of lines is equal to the perimeter L of the set K (Santalo (1976) p. 30).

In section 2.2 we will deal with the line segments of equal length. For this case we need the measure of the set of line segments which intersect a convex set with area F_0 and perimeter L_0. It is known (see Santalo (1976)) that this measure is equal to the sum $\pi F_0 + lL_0$, where l is the length of a linear segment.

If a line g intersects a figure D on the plane we write σ for the length of the intersection. According to the definition of the density of lines the following formula holds:

$$\int_{g \cap D \neq \emptyset} \sigma dG = \pi F, \qquad (2)$$

where F is the area of D and the integral is taken over all lines that intersect the figure D.

If we "normalize" the left-hand side of equation (2) by the measure of all lines that satisfy a condition A we obtain the conditional expectation $E(\sigma \mid A)$ of σ given that g satisfies A. In particular, if the condition A means that g intersects a convex figure K, or another convex figure K_1 containing K, the relation (2) implies the following two formulas for the expectation of the length σ of a random chord of K. First, the expectation over the set of lines that intersect K itself is

$$E(\sigma|K) = \frac{\pi F}{L}. \qquad (3)$$

We will write $E(\sigma)$ for $E(\sigma|K)$ below.

Second, if the figure K lies inside another convex set K_1, the conditional expectation of σ given that g intersects K_1 can be calculated using the formula

$$E(\sigma \mid K_1) = \frac{\pi F}{L_1}, \qquad (3')$$

where F is the area of K and L_1 is the perimeter of K_1.

We will also need the definition of breadth of a figure K:

The length of an orthogonal projection of a convex set K *to a line parallel to the direction* φ *is called the breadth of* K *in direction* φ *and we write* $\Delta(\varphi)$ *for it.*

The minimal (over φ) breadth is called the *width* of K and we write W to denote this. The following formula can be easily proved for a convex set K:

$$\mathrm{E}(\sigma) \leq \frac{\pi W}{2}, \tag{4}$$

where $\mathrm{E}(\sigma)$ is defined as above.

By a classic result of integral geometry about rectifiable curves C we know that the length L of C can be expressed as an integral of the number of intersection points $n(g, C)$ between this curve and a line g, i.e.

$$\int_G n(g, C)dG = 2L, \tag{5}$$

where the integral is taken over the set of all lines on the plane.

For line segments l of equal length we have

$$\int_{l \cap C \neq \emptyset} n(l, C)d\mathcal{L} = 2lL, \tag{5'}$$

where $\mathcal{L}$ is the set of line segments on the plane, L and l are the lengths of C and a line segment, respectively.

Equality (5) implies for the rectifiable curve C, which is located inside a closed convex curve C_1 with length L_1, that the expected number of intersection points between C and a line g over all lines that intersect C_1 is

$$\mathrm{E}(n(g, C)|C_1) = \frac{2L}{L_1}. \tag{6}$$

These nice formulae (3), (3') and (6) given above can be applied to estimate the area of figures or the length of curves by estimating the expectations on the left-hand side. The standard estimator of expectation is arithmetic mean. By the strong law of large numbers such an estimator is strongly consistent, i.e. with probability 1 the arithmetic mean converges to the expectation as the sample size tends to infinity.

However, in many cases it is necessary to make decisions (based on the same sample) on infinite sets of figures with unknown parameters and in such cases we need uniform almost sure convergence results. For example, suppose we want to identify the figure D^* with maximum area in some infinite class of figures. Then without uniform convergence property it may happen that the sequence of figures D_n^* with empirically maximum area will infinitely often contain the elements different from D^*.

In this paper we observe a large set of figures with unknown areas and we need uniform consistency to make decisions on the whole set of figures. We construct USC estimators using the methods of the theory of empirical processes.

2. Uniformly and strongly consistent estimator of the area of a planar figure obtained by using the direct approximation method

We will use the direct approximation method which is based on the two following theorems proved in Pollard (1984).

Let $\mathcal{F}$ be a class of measurable functions on a set S with a σ-field that carries a probability measure P. The empirical measure P_n is constructed by sampling from P. Assume $E \mid f \mid < \infty$ for each f in $\mathcal{F}$ and write $E_n f$ for the expectation with respect to P_n. In the context of section 1 $E_n f$ can be regarded as the arithmetic mean of the intersection points or of the length of a chord. The following theorems hold.

THEOREM 2.1. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathcal{F}_\varepsilon$ containing lower and upper approximations to each f in $\mathcal{F}$, such that $f_{\varepsilon,L} \le f \le f_{\varepsilon,U}$ and $E(f_{\varepsilon,U} - f_{\varepsilon,L}) < \varepsilon$. Then*

$$\sup_{\mathcal{F}} \mid E_n f - Ef \mid \to 0 \ \ a.s., \quad n \to \infty.$$

◇

THEOREM 2.2. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathcal{F}_\varepsilon$ of functions such that for each f in $\mathcal{F}$ there exists an f_ε in $\mathcal{F}_\varepsilon$ satisfying $f_\varepsilon \le f$ and $E(f_\varepsilon) \ge Ef - \varepsilon$. Then*

$$\liminf_n \inf_{\mathcal{F}} (E_n f - Ef) \ge 0 \ \ \textit{almost surely.}$$

◇

2.1. The basic convergence theorem. Let us consider a class $\mathcal{D}$ of uniformly bounded figures D on the plane. *

* This restriction of uniform boundedness has almost no importance from practical point of view, but it allows us to use the direct approximation method.

Let $G = \{g_i : 1 \leq i < \infty\}$ be independent uniformly distributed random lines on the plane defined by (1). The length of the intersection of a line g_i with a figure D is denoted by $\sigma_i = \sigma(g_i, \mathrm{D})$ and the number of intersection points between a line g_i and the boundary $\partial\mathrm{D}$ of D is denoted by $n(g_i, \partial\mathrm{D})$.

Our purpose is to prove that the empirical mean $\frac{1}{n}\sum_{i=1}^{n}\sigma_i$ converges to the expectation $\mathrm{E}(\sigma_1)$ with probability 1 *uniformly* over the class $\mathcal{D}$ as the number of lines increases indefinitely. The next theorem gives sufficient conditions for such convergence.

THEOREM 2.3. *Let $G = \{g_i : 1 \leq i < \infty\}$ be independent uniformly distributed random lines and $\mathcal{D}$ be a class of uniformly bounded figures* D *in* R^2. *Let every* D *from $\mathcal{D}$ have a rectifiable boundary. If* $\sup_{g_i \in G} \sup_{\mathrm{D} \in \mathcal{D}} n(g_i, \partial\mathrm{D}) \leq N < \infty$, *then*

$$\sup_{\mathrm{D}\in\mathcal{D}} \left| \frac{1}{n}\sum_{i=1}^{n}\sigma_i - \mathrm{E}(\sigma_1) \right| \to 0 \quad a.s., \quad n \to \infty.$$

PROOF

To apply the direct approximation method first note that every line g, defined by parameters p and φ, can also be regarded as a point $X = (p, \varphi)$ on the space $\mathrm{R} \times [0, 2\pi)$.

For each $\mathrm{D} \in \mathcal{D}$ we can define a function $f_{\mathrm{D}} : \mathrm{R} \times [0, 2\pi) \to \mathrm{R}$ as the length of the intersection between the line g and figure D, i.e. $f_{\mathrm{D}}(X) = \sigma(g, \mathrm{D})$. Write $\mathcal{F}_{\mathcal{D}}$ for the class of such functions f_{D}.

Now the statement of the current theorem can be written in the following form:

$$\sup_{f_{\mathrm{D}} \in \mathcal{F}_{\mathcal{D}}} \left| \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{D}}(X_i) - \mathrm{E}(f_{\mathrm{D}}(X_1)) \right| \to 0 \quad a.s. \tag{7}$$

By Theorem 2.1 it is sufficient to show that for each $\varepsilon > 0$ there exists a finite class $\mathcal{F}_{\mathcal{D}\varepsilon}$ of functions containing lower and upper approximations to each element of $\mathcal{F}_{\mathcal{D}}$ with uniformly small errors of approximation. As we have one-to-one correspondence between the classes $\mathcal{D}$ and $\mathcal{F}_{\mathcal{D}}$ we will first find such approximations to the elements of the class $\mathcal{D}$.

By assumption the class $\mathcal{D}$ is uniformly bounded, so without any loss of generality we can regard the elements of $\mathcal{D}$ be located in a finite square with a side M. For each $\varepsilon > 0$ this $\mathrm{M}\times\mathrm{M}$ square can be divided into smaller equal squares with side length $m = m(\varepsilon)$.

Let $\mathcal{R}(m)$ be the set of all possible figures made of these smaller $m \times m$ squares. It is clear that for each $\varepsilon > 0$ $\mathcal{R}(m)$ is finite and $card\mathcal{R}(m) = 2^{\left(\frac{M}{m}\right)^2}$.

Now, for each D from $\mathcal{D}$ we can find two corresponding figures from $\mathcal{R}(m)$ in such a way that one of them, say $D_{m,i}$, would be the biggest element of $\mathcal{R}(m)$ which is contained in D , and the other, say $D_{m,o}$, would be the smallest element of $\mathcal{R}(m)$ which contains D .

Write $f_r(X) = \sigma(g, r)$ for the length of the intersection of a line g with the element r of $\mathcal{R}(m)$ and let $\mathcal{F}_{\mathcal{R}(m)}$ be the class of such functions, i.e.

$$\mathcal{F}_{\mathcal{R}(m)} = \{f_r(X) | r \in \mathcal{R}(m)\}.$$

Define now $\mathcal{F}_{\mathcal{D}\varepsilon} := \mathcal{F}_{\mathcal{R}(m)}$, $f_{\mathrm{D}}^{\varepsilon,i} := f_{\mathrm{D}_{m,i}}$, $f_{\mathrm{D}}^{\varepsilon,o} := f_{\mathrm{D}_{m,o}}$.

It is clear that the class $\mathcal{F}_{\mathcal{D}\varepsilon}$ is finite and for each $f_{\mathrm{D}} \in \mathcal{F}_{\mathcal{D}}$ it contains two functions $f_{\mathrm{D}}^{\varepsilon,i}$ and $f_{\mathrm{D}}^{\varepsilon,o}$ such that

$$f_{\mathrm{D}}^{\varepsilon,i} \leq f_{\mathrm{D}} \leq f_{\mathrm{D}}^{\varepsilon,o}.$$

To see that the second condition of Theorem 2.1 is also satisfied we have to show that for every $\varepsilon > 0$

$$\mathrm{E}(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i}) < \varepsilon \tag{8}$$

and we proceed as follows.

By assumption every D from $\mathcal{D}$ has rectifiable boundary and the number of intersection points between the boundary of D and a line g is finite. Uniform boundedness guarantees us that for each D from $\mathcal{D}$ $n(g, \partial \mathrm{D}) = 0$ outside the M×M square. Let us write L_{D} for the length of the boundary of D. Now, as a square is convex, we have from (5) the following:

$$2L_{\mathrm{D}} = \int_G n(g, \partial \mathrm{D}) dG = \int_{g \cap (\mathrm{M} \times \mathrm{M}) \neq \emptyset} n(g, \partial \mathrm{D}) dG \leq$$

$$\leq N \cdot \int_{g \cap (\mathrm{M} \times \mathrm{M}) \neq \emptyset} dG = N \cdot 4\mathrm{M}.$$

Let us set $L = 2N \cdot \mathrm{M}$. So, as $L < \infty$, we have that every D from $\mathcal{D}$ has boundary with some finite length L_{D} , $L_{\mathrm{D}} \leq L$.

Recall that the M×M square is divided into $m \times m$ squares. It is easy to see that a segment of a planar closed curve of length m cannot intersect more than 4 squares with side m. Hence the boundary of D with length L_{D} does not intersect more than $\frac{4L_{\mathrm{D}}}{m}$ squares and, therefore, the total area $F_{\varepsilon,\mathrm{D}}$ of such squares is bounded by $4L_{\mathrm{D}}m \leq 4Lm$.

Now, if for a given $\varepsilon > 0$ we choose m so that $m < \dfrac{\varepsilon^2}{4\pi L}$, we get

$$F_{\varepsilon,\mathrm{D}} < \frac{\varepsilon^2}{\pi}. \tag{9}$$

Let us now divide the elements of $\mathcal{D}$ into two subclasses according to the widths of their convex hulls. With a slight abuse in the notation we will talk about the breadth and the width of D meaning actually the breadth and the width of its convex hull. This helps us to overcome the difficulties with nonconvex figures. For each $\varepsilon > 0$ we fix so-called critical value of width

$$W_\varepsilon = \frac{\varepsilon}{\pi} \tag{10}$$

and observe separately the "thick" figures with width W strictly larger than W_ε, and the "thin" figures with width $W \leq W_\varepsilon$. Let us first consider the case $W > W_\varepsilon$.

Using formula (2) we first find the bound to the expectation having φ fixed:

$$\mathrm{E}_\varphi(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i}) = \int\limits_{g\cap\mathrm{D}\neq\emptyset} \frac{(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i})}{\Delta(\varphi)} dp \leq \frac{F_{\varepsilon,\mathrm{D}}}{\Delta(\varphi)}, \quad \varphi \in [0,\pi].$$

As $\Delta(\varphi) \geq W$ for each φ, then taking expectation over φ gives us

$$\mathrm{E}(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i}) \leq \int\limits_0^\pi \frac{F_{\varepsilon,\mathrm{D}}}{\Delta(\varphi)} \cdot \frac{1}{\pi} d\varphi \leq \frac{F_{\varepsilon,\mathrm{D}}}{W \cdot \pi} \int\limits_0^\pi d\varphi = \frac{F_{\varepsilon,\mathrm{D}}}{W} < \frac{F_{\varepsilon,\mathrm{D}}}{W_\varepsilon} < \varepsilon,$$

where the last inequality follows from (9) and (10). So (8) is satisfied for "thick" figures.

Now consider figures with convex hulls of width $W \leq W_\varepsilon$. Let us write f_{ch} for the length of a random chord of a convex hull. Then, by (4), we have

$$\mathrm{E}(f_{ch}) \leq \frac{\pi W}{2} \leq \frac{\pi W_\varepsilon}{2} < \varepsilon.$$

From the other side we have simple inequalities

$$\mathrm{E}(f_{\mathrm{D}} - f_{\mathrm{D}}^{\varepsilon,i}) \leq \mathrm{E}(f_{\mathrm{D}}) \leq \mathrm{E}(f_{ch}).$$

Therefore we obtain $\mathrm{E}(f_{\mathrm{D}} - f_{\mathrm{D}}^{\varepsilon,i}) < \varepsilon$.

With this we have shown that there exists a class of functions that contains at least one-sided (for "thick" figures even two-sided) approximation(s) to each element of $\mathcal{F}_{\mathcal{D}}$. Applying Theorem 2.2 we get the convergence (7).

q.e.d.

The next corollary is a direct application of Theorem 2.3 to the set of convex figures. Since every convex curve is rectifiable and every line can have at most two intersection points with the boundary of a convex figure the following theorem is valid.

COROLLARY 2.1. *Consider independent uniformly distributed random lines* $G = \{g_i : 1 \leq i < \infty\}$ *and any class* $\mathcal{K}$ *of uniformly bounded convex figures* K *on the plane. Then*

$$\sup_{\mathrm{K}\in\mathcal{K}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i - \mathrm{E}(\sigma_1)\right| \to 0 \quad a.s., \quad n \to \infty.$$

◇

2.2. USC estimator for the area of a planar figure. Let us now apply the direct approximation method to derive USC estimator of the areas of figures on the plane. We will generalize the "line intercept sampling" method which is used in plant ecology in estimating the proportion of plants in some fixed sampling area (Pielou (1985)). The proportion of a plant species is defined as the share of the total fixed sampling area covered by that plant species. The main difference between this method and the approach we have used till now is that here the line segments with finite length l are used instead of straight lines with infinite length. Nevertheless, the idea of estimation remains the same – the length of the total intercept is measured. More precisely, a line segment of length l is thrown onto the sampling area at random and then the estimate of the proportion V of the total area covered by some plant species is given by the estimator

$$V^* = \frac{1}{l}\sum_{j=1}^{J} l_j, \tag{11}$$

where l_j is the length of the j-th plant patch intercept and J is the number of patches cut by the sampling line segment. It is known (see Kendall, Moran (1972)) p. 79) that this estimator is unbiased, i.e.

$$\mathrm{E}(V^*) = V. \tag{12}$$

Theorem 2.4 gives the generalization of this method in the sense that instead of one random line segment we use the set of random line segments and therefore unbiasedness can be strengthened to uniform consistency. The proof parallels the way we proceeded while proving Theorem 2.3.

THEOREM 2.4. *Consider a class $\mathcal{D}$ of uniformly bounded figures* D *with rectifiable boundaries and n independent uniformly distributed random line segments $l^{(1)}, \ldots, l^{(n)}$ of length l in* R^2. *Let $J_{\mathrm{D},i}$ be the number of connected subregions of* D *intersected by a line segment $l^{(i)}$, $i = 1, \ldots, n$, and let $\sum_{j=1}^{J_{\mathrm{D},i}} l_j^{(i)}$ be the total length of the intersection of a line segment $l^{(i)}$ with figure* D, $i = 1, \ldots, n$. *If* $\sup_{\mathrm{D}\in\mathcal{D}} \sup_i J_{\mathrm{D},i} \leq J < \infty$, *then*

$$V_{\mathrm{D}}^* = \frac{1}{nl} \sum_{i=1}^{n} \sum_{j=1}^{J_{\mathrm{D},i}} l_j^{(i)} \tag{13}$$

is USC estimator of the proportion of the area of D.

PROOF

For the sake of simplicity let the sampling area be M×M square, where M is chosen sufficiently large for all the elements of the class $\mathcal{D}$ to be bounded by that square.

Every line segment on the plane is determined by the coordinates of its center (x,y) and by the angle φ between it and the x-axis. We can define each line segment as a point $X = (x, y; \varphi)$ on the space $\mathrm{R}^2 \times [0, \pi)$. For each D from $\mathcal{D}$ we can define a function $f_{\mathrm{D}} : \mathrm{R}^2 \times [0, \pi) \to \mathrm{R}$,

$$f_{\mathrm{D}}(X) = \frac{1}{l} \sum_{j=1}^{J_{\mathrm{D}}} l_j, \tag{14}$$

where l_j is the j-th intercept of the line segment X with figure D and J_{D} is the number of intercepts.

Having in mind (11) and (12) we get from (14) that, for each D in $\mathcal{D}$,

$$\mathrm{E}(f_{\mathrm{D}}(X)) = V_{\mathrm{D}} \,. \tag{15}$$

Notice that $\frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{D}}(X_i)$ is equal to (13). As we consider independent line segments with the same distribution, the summands $f_{\mathrm{D}}(X_i)$ are i.i.d random variables and by the strong law of large numbers we have

$$\frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{D}}(X_i) \to \mathrm{E}(f_{\mathrm{D}}(X_1)) \ \ a.s., \ \ n \to \infty. \tag{16}$$

So we are left to show that the convergence (16) is uniform over the class $\mathcal{D}$.

Let us fix an $\varepsilon > 0$. Exactly in the same manner as in the proof of Theorem 2.3 we can find inner and outer approximations – $\mathrm{D}_{\varepsilon,i}$ and $\mathrm{D}_{\varepsilon,o}$, respectively, – made of finite number (depending on ε) of small squares to each element of $\mathcal{D}$. This gives us also approximations for the functions (14), say $f_{\mathrm{D}}^{\varepsilon,i}$ and $f_{\mathrm{D}}^{\varepsilon,o}$, respectively, which satisfy the conditions

$$f_{\mathrm{D}}^{\varepsilon,i} \le f_{\mathrm{D}} \ \le f_{\mathrm{D}}^{\varepsilon,o}$$

and

$$\mathrm{E}(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i}) < \varepsilon. \tag{17}$$

Indeed, the first condition is trivial and the second comes from the definition of area proportion and from the fact that the area $F_{\varepsilon,\mathrm{D}}$ between the inner and outer approximations can be made arbitrary small: using (15) we can first write

$$\mathrm{E}(f_{\mathrm{D}}^{\varepsilon,o} - f_{\mathrm{D}}^{\varepsilon,i}) = \mathrm{E}(f_{\mathrm{D}}^{\varepsilon,o}) - \mathrm{E}(f_{\mathrm{D}}^{\varepsilon,i}) = V_{\mathrm{D}_{\varepsilon,o}} - V_{\mathrm{D}_{\varepsilon,i}} = \frac{F_{\varepsilon,\mathrm{D}}}{\mathrm{M}^2},$$

where M^2 is the area of the whole sampling area. Notice that for each D we have $n(l^{(i)}, \partial\mathrm{D}) \le 2J_{\mathrm{D},i}$. Now, if we write L_{D} for the length of the perimeter of D we have by (5′) that

$$2lL_{\mathrm{D}} = \int\limits_{l\cap\partial\mathrm{D}\neq\emptyset} n(l^{(i)}, \partial\mathrm{D})d\mathcal{L} \le \int\limits_{l\cap\mathrm{D}\neq\emptyset} n(l^{(i)}, \partial\mathrm{D})d\mathcal{L} \le \int\limits_{l\cap\mathrm{D}\neq\emptyset} 2J_{\mathrm{D},i}d\mathcal{L} \le$$

$$\le 2J \cdot \int\limits_{l\cap\mathrm{D}\neq\emptyset} d\mathcal{L} \le 2J \cdot \int\limits_{l\cap(\mathrm{M}\times\mathrm{M})\neq\emptyset} d\mathcal{L} = 2J(\pi\mathrm{M}^2 + 4l\mathrm{M}) < \infty.$$

So, each D from $\mathcal{D}$ has boundary with finite length. But this, as we have already seen in the proof of Theorem 2.3, is sufficient for

that that for a given $\varepsilon > 0$ $F_{\varepsilon,\mathrm{D}}$ can be made arbitrary small. So, as soon as the condition $F_{\varepsilon,\mathrm{D}} < \mathrm{M}^2 \cdot \varepsilon$ is fulfilled (17) holds. Now, by Theorem 2.1 (16) holds uniformly over $\mathcal{D}$. This completes the proof.

q.e.d.

COROLLARY 2.2. *Consider a class $\mathcal{D}$ of uniformly bounded (by a set* K *of area F_K) figures* D *having rectifiable boundaries. Let $l^{(i)}$, $i = 1, \ldots, n$ be independent uniformly distributed random line segments of length l in* R^2. *Let $J_{\mathrm{D},i}$ be the number of connected subregions of* D *intersected by a line segment $l^{(i)}$, and let $\sum_{j=1}^{J_{\mathrm{D},i}} l_j^{(i)}$ be the total length of the intersection of a line segment $l^{(i)}$ with figure* D, $i = 1, \ldots, n$. *If* $\sup_{\mathrm{D}\in\mathcal{D}} \sup_i J_{\mathrm{D},i} < \infty$, *then*

$$F_{\mathrm{D}}^* = \frac{F_{\mathrm{K}}}{nl} \sum_{i=1}^{n} \sum_{j=1}^{J_{\mathrm{D},i}} l_j^{(i)}$$

is USC estimator of the area of D.

PROOF

Take K in the role of the sampling area, apply Theorem 2.4 and use the relationship $F_{\mathrm{D}}^* = F_{\mathrm{K}} \cdot V_{\mathrm{D}}^*$.

q.e.d.

Notice that the shape of the set in which the observable class is located, has no importance. So in practical cases we can always take it such that measuring its area would not be a problem (square, circle etc).

Acknowledgement. I would like to thank my supervisor Assoc. Prof. Kalev Pärna for useful advice and suggestions that he gave me while I was writing my master thesis where part of these results were obtained.

References.

Kendall, M. and Moran, P. (1972) Geometrical Probabilities. Moscow: "Nauka", (in Russian).

Pielou, E.C. (1985) "Line intercept sampling." – In: Encyclopedia of Statistical Sciences. Edited by Kotz, S., Johnson, N. L., Vol 5, p 70, J. Wiley & Sons, New York.

Pollard, D. (1984) Convergence of Stochastic Processes. Springer Series in Statistics, New York.

Santalo, L. A. (1976) Integral Geometry and Geometric Probability. Vol. 1 of Encyclopedia of Mathematics and Its Applications. Addison-Wesley Publ. Comp.

Received November, 1993.

Näide ühtlaselt tugevast suurte arvude seadusest juhuslike sirgete korral.

Tarmo Koll

Kokkuvõte

Käesolevas artiklis vaadeldakse juhuslike sirgete lõikeid mingist loenduvast tõkestatud hulgast pärinevate tasandiliste kujunditega. Tõestatakse teoreem sellise juhusliku lõike pikkuse aritmeetilise keskmise ühtlasest koondumisest peaaegu kindlasti tema keskväärtuseks. Selleks kasutatakse empiiriliste protsesside teooriast tuntud otsese lähendamise meetodit (Pollard (1984)). Seejärel rakendatakse sama tõestusideed tasandilise kujundi pindalale ühtlaselt ja tugevalt mõjusa hinnangu saamiseks juhuslike sirglõikude abil. Selleks üldistatakse taimeökoloogiast tuntud "lõiguga tabamise" meetodit. Saadud tulemus lubab teha statistilisi järeldusi ka üle loenduva hulga tasandiliste kujundite.

Acta et Commentationes Universitatis Tartuensis, 968 ,1994, pp. 17-27.

A Note on Patterned Matrices with Applications in Multivariate Analysis

Tõnu Kollo

Summary. In the paper a notion of the patterned matrix is introduced for a matrix A independently of functional relations between elements of the matrix. The transformation of a matrix into a patterned matrix is realized by a transition matrix. An explicit expression of the transition matrix which has the unique inverse transformation under certain conditions has been presented. Introduced notions are used in two applications: firstly it has been shown that in deriving asymptotic covariance matrices for asymptotically normal statistics one can neglect the functional relations between elements of the considered multivariate statistic and the explicit expressions of the transition matrices have been found to select nonrepeated multivariate moments.

Key words: *patterned matrix, transition matrix, matrix derivative, asymptotic distribution, higher order moments.*

1. Introduction. Patterned matrices have been studied in recent years in several papers (Tracy, Jinadasa (1987, 1988); Wiens (1985), for example). Usually a matrix A is looked as a patterned matrix if there is some relationship between its elements e.g. symmetric matrices. In this paper we do not intend to connect the notion of patterned matrix with any kind of relationship among the elements of A. By definition (DEFINITION 1) we are talking about a patterned matrix A(K), if some element(s) or certain part of the original matrix A has been excluded, "cut off", from A. As a special case, the notion of amputated matrix has been introduced by Parring (1980) for examining matrices, from where certain columns or rows where excluded. In matrix differentiation problems patterned matrices are used for eliminating constants and

repeated elements from matrices. This includes the possibility to get any patterned matrix in our sense, but the difference is that a patterned matrix must not consist of all non-repeated variables or may include part of repeated variables also.

In section 2 we are going to introduce basic notions and notations on patterned matrices, which we need later in applications. In section 3 the problem of differentiating of a matrix function by a patterned matrix is examined and in section 4 patterned matrices are used in working with higher order multivariate moments.

2. Patterned matrices. Let A be a $p \times q$-matrix and K the set of pairs of indices:

$$\mathrm{K} = \{(i,j): i \in \mathrm{I_K}; j \in \mathrm{J_K}; \mathrm{I_K} \subset \{1,\ldots,p\}; \mathrm{J_K} \subset \{1,\ldots,q\}\}.$$

DEFINITION 1. A(K) *is a patterned matrix, if* A(K) *consists of elements* a_{kl} *of* A, $(k,l) \in \mathrm{K}$.

Nel (1980) summarized the results on patterned matrices till 1980 and for many notions and notations we will refer to him.

Following Anderson (1958) we use double indices for indicating elements of block-matrices: the element of a block-matrix A in j-th row of i-th row of blocks and l-th column of k-th column of blocks is denoted by $a_{(i,j)(k,l)}$ (or $(\mathrm{A})_{(i,j)(k,l)}$). It means that in the index $(i,j)(k,l)$ i and k stay for block-indices, j and l are row and column indices correspondingly. If A consists of one column (one row) of blocks we use the notation $a_{(i,j)k}$ $(a_{i(k,l)})$ for the element of A.

Let A be an $p \times q$-matrix and A_j its j-th column, then *vec*A is the pq-vector

$$vec\mathrm{A} = \begin{pmatrix} \mathrm{A}_1 \\ \vdots \\ \mathrm{A}_q \end{pmatrix}.$$

When the elements of A(K) are collected into one column by columns of A in natural order, we get a vector with dimensionality k, where k is the number of pairs in K. Let us denote this vector by *vec*A(K). Nel (1980) uses the notation *vec*pA for the *vec*A from which all the constants and repeated elements are eliminated. Henderson, Searle (1979) are using *vech*A to indicate the vector obtained from the lower triangular part of symmetric A, Traat (1986) and McCullogh (1982) use this notation for the upper triangle, Magnus, Neudecker (1978) denote the last vector by $v(\mathrm{A})$.

To transform vecA into vecA(K) we have to introduce $k \times pq$-transition matrix T(K):

$$vec\mathrm{A}(\mathrm{K}) = \mathrm{T}(\mathrm{K})vec\mathrm{A}, \tag{1}$$

where we look T(K) as a block-row matrix, consisting of $k \times p$-blocks.

When the f-th coordinate of vecA(K) is the element a_{gh} of the original matrix A we use the following indexation: $(vec\mathrm{A}(\mathrm{K}))_{f(gh)}$. Then the simpliest way of defining T(K) is via equality (2):

$$(\mathrm{T}(\mathrm{K}))_{f(gh)(i,j)} = \begin{cases} 1, & i = h, \ j = g; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

It is not the only way of defining transformation (1) if there exist functional relations between elements of A. Later we will use another way of defining T(K) to get a more convenient expression from point of view of considering inverse transformations.

EXAMPLE.

Let R be a 3×3-correlation matrix:

$$\mathrm{R} = \begin{pmatrix} 1 & r_{12} & 0 \\ r_{21} & 1 & r_{23} \\ 0 & r_{32} & 1 \end{pmatrix}$$

and we assume we want to select as R(K) all the elements not equal to 1 or 0. Then K=$\{(1,2);(2,1);(2,3);(3,2)\}$ and

$$vec\mathrm{R}(\mathrm{K}) = \begin{pmatrix} r_{21} \\ r_{12} \\ r_{32} \\ r_{23} \end{pmatrix} = \mathrm{T}(\mathrm{K})vec\mathrm{R},$$

where from formula (2)

$$\mathrm{T}(\mathrm{K}) = \left(\begin{array}{ccc|ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right).$$

If we have additional information about the structure and elements of A, we are able to define T(K) in such a way that we can consider also the inverse operation to construct vecA from vecA(K) .

DEFINITION 2. *We call a $k \times pq$-block-row matrix* T(K) *the transition matrix for a patterned matrix* A(K) *if* T(K) *consists of $k \times p$-blocks and for $f(gh) \in \{1, \ldots, k\}$ we have:*

$$\mathrm{T(K)}_{f(gh)(i,j)} = \begin{cases} \frac{1}{m(gh)}, & a_{ij} = a_{gh}, \\ & m(gh) = \sum_{i=1}^{p} \sum_{j=1}^{q} \mathbf{1}_{\{|a_{ij}|=|a_{gh}|\}}; \\ -\frac{1}{m(g,h)}, & -a_{ij} = a_{gh}; \\ 0, & \textit{otherwise.} \end{cases}$$

In the definition the indicator function for the absolute values of two elements is defined as usual:

$$\mathbf{1}_{\{a=b\}} = \begin{cases} 1, & a = b; \\ 0, & a \neq b. \end{cases}$$

EXAMPLE (*continued*).

Let us find the transition matrix in the case of matrix R for the same set K with the additional information, that R is symmetric. Then

$$vec\mathrm{R(K)} = \begin{pmatrix} r_{21} \\ r_{12} \\ r_{32} \\ r_{23} \end{pmatrix} = \mathrm{T(K)}vec\mathrm{R},$$

where from Definition 2

$$\mathrm{T(K)} = \left(\begin{array}{ccc|ccc|ccc} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right).$$

Let us assume, that we want to have in R(K′) all different correlations $r_{ij} \neq 0, 1$. Then $\mathrm{K}' = \{(1,2);(2,3)\}$ and

$$\mathrm{T(K')} = \left(\begin{array}{ccc|ccc|ccc} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right).$$

Remark that for $\mathrm{K}'' = \{(2,1);(3,2)\}$ the transition matrix T(K″) will be the same, so the transformation does not depend on which equal elements we have fixed in the set K.

To define the inverse transformation which realizes the mapping

$$vec\mathrm{A(K)} \to vec\mathrm{A}$$

we need all the different nonzero elements of A being presented in *vec*A(K). For example, we cannot reconstruct *vec*R by *vec*R(K) or by *vec*R(K′) in our example because we have no constant 1 in *vec*R(K) .

Let us assume now that *vec*A(K) includes all the elements excluding zero. Then the inverse transition matrix P(K) is defined implicitly by the equality (3)

$$\mathit{vec}\mathrm{A} = \mathrm{P(K)}\mathit{vec}\mathrm{A(K)}. \tag{3}$$

Nel (1980) proved that

$$\mathrm{P(K)} = \mathrm{T(K)}^{+}; \quad \mathrm{T(K)} = \mathrm{P(K)}^{+},$$

if T(K) is the transition matrix to eliminate constants and repeated elements from matrix A (here $\mathrm{T(K)}^{+}$ denotes the Moore-Penrose inverse of T(K)). Nel's proof extends to our case straightforwardly, so we get

$$\mathit{vec}\mathrm{A} = \mathrm{T(K)}^{+}\mathit{vec}\mathrm{A(K)}, \tag{4}$$

if all non-zero elements of A are presented in A(K) .

EXAMPLE (*continued*).

Let now $\mathrm{K} = \{(i,j): i = 1,2,3;\ j = 2\}$. Then we have all non-zero elements in R(K) and

$$\mathrm{T(K)} = \left(\begin{array}{ccc|ccc|ccc} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}\right).$$

For the Moore-Penrose inverse we get the equality

$$\mathrm{T(K)}^{+} = \left(\begin{array}{c} 0\ 1\ 0 \\ 1\ 0\ 0 \\ 0\ 0\ 0 \\ - \ - \ - \\ 1\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \\ - \ - \ - \\ 0\ 0\ 0 \\ 0\ 0\ 1 \\ 0\ 1\ 0 \end{array}\right).$$

It is easy to check that all the four defining properties of the Moore-Penrose inverse (see Rao (1965), for example) are satisfied and

$$\mathrm{T(K)}^{+}\begin{pmatrix} r_{12} \\ 1 \\ r_{32} \end{pmatrix} = vec\mathrm{R}.$$

3. An application in asymptotic distribution theory. Asymptotic normality can be established for a very wide class of statistics in multivariate analysis. All continuously differentiable functions of the sample covariance matrix or the sample mean are asymptotically normal, for example. If matrix X is a random matrix and it is asymptotically normal, then asymptotic normality can be established for $\mathrm{Y} = \mathrm{Y(X)}$ also and for the asymptotic covariance matrix DY_{as} of Y we have the following expression:

$$\mathrm{DY}_{as} = \frac{d\mathrm{Y}}{dvec'\mathrm{X(K)}}\bigg|_{\mathrm{X=EX}} \mathrm{D}vec\mathrm{X(K)}\left(\frac{d\mathrm{Y}}{dvec'\mathrm{X(K)}}\right)'_{\mathrm{X=EX}}, \quad (5)$$

where $\dfrac{d\mathrm{Y}}{dvec'\mathrm{X(K)}}$ is a suitably defined matrix derivative (Magnus, Neudecker (1988), or Kollo (1991)), and X(K) is obtained from X by eliminating constants and repeated elements. It is rather troublesome to use $vec\mathrm{X(K)}$ in differentiating because of loosing usual matrix operations and properties. Would be much more convenient to use $vec\mathrm{X}$ directly. The following lemma gives us bases for that.

LEMMA. *Let* A *be an arbitrary* $p \times q$*-matrix,* $vec\mathrm{A(K)}$ *the* k*-vector consisting of all distinct coordinates of* $vec\mathrm{A}$ *and* T(K) *its transition matrix. Then for an arbitrary* $k \times k$*-matrix* M(k)

$$\mathrm{T(K)}^{+}\mathrm{M(k)}(\mathrm{T(K)}^{+})' = \mathrm{M},$$

where in the $pq \times pq$*-matrix* M *i-th and j-th rows and columns are equal to each other, if in* vecA *i-th and j-th coordinates are equal; i-th row and column of* M *are zeros if i-th coordinate of* vecA *is zero.*

PROOF.

The result follows from the equality (4). In every column of the product $[\mathrm{T(K)}^{+}\mathrm{M(k)}]$ i-th and j-th coordinates are equal, if the same coordinates are equal in vecA. Also in the product $[\mathrm{T(K)}^{+}\mathrm{M(k)}](\mathrm{T(K)}^{+})'$ in every row i-th and j-th coordinates are equal if these coordinates are equal in vecA.

From our lemma we get the following result.

COROLLARY. *The asymptotic covariance matrix (5) of a random matrix* $Y=Y(X)$ *can be found as*

$$DY_{as} = \left.\frac{dY}{dX}\right|_{X=EX} D\,vec X \left(\frac{dY}{dX}\right)'_{X=EX},$$

if in $vecX(K)$ *all different variable elements of* X *are presented.*

PROOF.

By properties of the matrix derivative (chain rule)

$$\frac{dY}{d\,vec'X(K)} = \frac{dY}{dX}\frac{dX}{d\,vec'X(K)},$$

if all different variable elements of X are presented in X(K). Differentiating equality (4) we obtain

$$\frac{d\,vec X}{d\,vec'X(K)} = T(K)^{+}\frac{d\,vec X(K)}{d\,vec'X(K)} = T(K)^{+}.$$

By definition of the matrix derivative we have

$$\frac{dX}{d\,vec'X(K)} = \frac{d\,vec X}{d\,vec'X(K)}$$

and then from the expression (5) we get

$$\begin{aligned} DY_{as} &= \left.\frac{dY}{d\,vec'X(K)}\right|_{X=EX} D\,vec X(K) \left(\frac{dY}{d\,vec'X(K)}\right)_{X=EX} = \\ &= \left.\frac{dY}{dX}\right|_{X=EX} T(K)^{+} D\,vec X(K) \left(T(K)^{+}\right)' \left(\frac{dY}{dX}\right)'_{X=EX} = \\ &= \left.\frac{dY}{dX}\right|_{X=EX} D\,vec X \left(\frac{dY}{dX}\right)'_{X=EX}. \end{aligned}$$

The last equality follows from the lemma.

It means that if we find matrix derivatives for calculating asymptotic covariance matrices we can forget about the relationship between elements of the matrix X and if $x_{ij} = const$, the corresponding column in the derivative matrix is a zero column $\left(\frac{\partial y_{kl}}{\partial x_{ij}} = 0 \text{ for any } y_{kl}\right)$.

4. Transition matrices for selecting multivariate moments. The problem of determing transition matrices for non-symmetric patterned matrices arises in multivariate analysis when finding multivariate higher order moments and cumulants as derivatives of a characteristic function. Then we are interested in selecting nonequal elements from the matrix of partial derivatives of k-th order. In fact we have the same problem when dealing with moments and cumulants of a random vector directly. In this case the k-th order moment $m_k(\mathrm{X})$ is defined through Kronecker product, for example

$$m_k(\mathrm{X}) = \mathrm{E}(\mathrm{X} \otimes \mathrm{X} \otimes \cdots \otimes \mathrm{X})$$

and we are interested in collecting all the different elements from the product.

In Kollo, von Rosen (1993) an algorithm has been proposed for this procedure. Here we are going to present the transition matrices realizing that algorithm. To describe the algorithm we need the following notions and notations.

Let $[i, j]$ stand for the number of combinations given by

$$[i, j] = \binom{i-1+j-1}{i-1} \quad i, j = 1, 2, \ldots \tag{6}$$

and

$$[0, j] = [i, 0] = 0.$$

As a basic notion we define the vectorizing operator $V^j(\mathrm{A})$.

DEFINITION 3. *For any matrix* $A : [j, n] \times n, \quad j = 1, 2, \ldots$

$$V^j(\mathrm{A}) = (a_{11}, a_{12}, \ldots, a_{[j,2]2}, a_{13}, \ldots, a_{[j,3]3}, \ldots, a_{1n}, \ldots, a_{[j,n]n})'.$$

In particular, for $j = 1, 2, 3$, we have

$$\begin{aligned}
V^1(\mathrm{A}) &= (a_{11}, a_{12}, \ldots, a_{1n})', \\
V^2(\mathrm{A}) &= (a_{11}, a_{12}, a_{22}, a_{13}, a_{23}, a_{33}, \ldots, a_{1n}, a_{2n}, \ldots, a_{nn})', \\
V^3(\mathrm{A}) &= (a_{11}, a_{12}, a_{22}, a_{32}, a_{13}, a_{23}, a_{33}, a_{43}, a_{53}, a_{63}, \ldots, a_{\frac{n(n+1)}{2}n})'.
\end{aligned}$$

It has been proved (Kollo, von Rosen (1993)) that all different mixed moments and cumulants of higher order can be found by vectorizing certain matrices A with the operator $V^j(\mathrm{A})$. For using this result in applications the transition matrices for the operators $V^j(\mathrm{A})$ are needed. In the following we present the structure of

these matrices. Because we are not interested in reconstructing matrices of Kronecker products from their nonrepeated elements, we will use the formula (2) in finding transition matrices. At first, let us write out the expressions of transition matrices for $j = 1, 2$.

If $j = 1$, A is $1 \times n$ matrix, corresponding to

$$\mathrm{K}^1 = \{(i, j): i = 1, j = 1, \ldots, n\}$$

and $V^1(\mathrm{A})$ is n-vector. So

$$\mathrm{T}(\mathrm{K}^1) = 1.$$

If $j = 2$, we get from the formula (6) that $(2, n) = n$ and A is $n \times n$-matrix. Then

$$\mathrm{K}^2 = \{(i, j): i \leq j, \quad i, j = 1, \ldots, n\}$$

and the transition matrix $\mathrm{T}(\mathrm{K}^2)$ for cutting out the upper triangle of A is $\left(\frac{1}{2}n(n+1) \times n^2\right)$-matrix. From (2) it comes out that $\mathrm{T}(\mathrm{K}^2)$ has the block-diagonal structure and the i-th diagonal block $[\mathrm{T}(\mathrm{K}^2)]_{ii}$ has the following form:

$$[\mathrm{T}(\mathrm{K}^2)]_{ii} = \begin{pmatrix} \mathrm{I}_{(1)} \\ \vdots \\ \mathrm{I}_{(i)} \end{pmatrix},$$

where $\mathrm{I}_{(k)}$ denotes the k-th row of the identity matrix I_n. In the general case the transition matrix $\mathrm{T}(\mathrm{K}^j)$ for the vector $V^j(\mathrm{A})$ has the similar structure. Matrix $\mathrm{T}(\mathrm{K}^j)$ is block-diagonal with diagonal blocks

$$[\mathrm{T}(\mathrm{K}^j)]_{ii} = \begin{pmatrix} \mathrm{I}_{(1)} \\ \mathrm{I}_{(2)} \\ \vdots \\ \mathrm{I}_{((j,i))} \end{pmatrix}, \quad i = 1, \ldots, n,$$

where $\mathrm{I}_{(k)}$ is the k-th row of $\mathrm{I}_{(j,n)}$, and (i, j) is defined by equality (6).

References.

Anderson, T.W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley.

Henderson, H.V., Searle, S.R. (1979) Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. Canad. J. Statist. **7**, 65-81.

Kollo, T. (1991) Matrix Derivative in Multivariate Statistics. Tartu University Press, Tartu (in Russian).

Kollo, T., von Rosen, D. (1993) Moments and Cumulants of Symmetric Matrices with Applications to the Wishart Distribution. U.U.D.M. Report 1993:10, Uppsala University, Department of Mathematics, Uppsala.

Magnus, J.R., Neudecker, H. (1978) The Elimination Matrix: Some Theorems and Applications. Report AE 2/78, Faculty of Actuarial Science and Econometrics, University of Amsterdam.

Magnus, J.R., Neudecker, H. (1988) Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley.

McCullogh, C.E. (1982) Symmetric matrix derivative with applications. J. Amer. Statist. Assoc.,**77**, 679-682.

Nel, D.G. (1980) On matrix differentiation in statistics. South African Statist. J. ,**14**, 137-193.

Parring, A.M. (1980) The asymptotic distribution of incomplete multiple correlation coefficients. Acta et Commentationes Universitatis Tartuensis, **541**, 18-26 (in Russian).

Rao, C.R. (1965) Linear Statistical Inference and Its Applications. Wiley, New York.

Traat, I. (1986) Matrix calculus for multivariate distributions. Acta et Commentationes Universitatis Tartuensis, **733**, 64-85.

Tracy, D.S., Jinadasa, K.G. (1987) Matrix derivatives of patterned matrices and statistical applications. New Perspectives in Theoretical and Applied Statistics. Edited by M.L. Puri, J.P. Vilaplan, W. Wertz, Wiley, 341-351.

Tracy, D.S., Jinadasa, K.G. (1988) Patterned matrix derivatives. Canad. J. Statist., **16**, 411-418.

Wiens, D.P. (1985) On some pattern-reduction matrices which appear in statistics. Linear Algebra and Its Applications, **67**, 233-258.

Received December, 1993

Kujundmaatriksitest koos rakendustega mitmemõõtmelises analüüsis

Tõnu Kollo

Kokkuvõte

Artiklis on sisse toodud kujundmaatriksi mõiste maatriksi A jaoks sõltumatult maatriksi A elementide vahelistest funktsionaalsetest seostest. Teisendus, mis seob kujundmaatriksit maatriksiga A, on antud tema üleminekumaatriksiga. On saadud tingimused üleminekumaatriksi pöördteisenduse olemasoluks ja ühesuseks ning leitud üleminekumaatriksi üldelemendi avaldis. Esitatud mõisteid kasutatakse kahe ülesande lahendamisel. Esiteks tõestatakse, et asümptootilise normaaljaotuse korral võib asümptootilise kovariatsioonimaatriksi leidmisel eirata statistiku elementide vahelisi funktsionaalseid seoseid. Teiseks leitakse üleminekumaatriksite avaldised tensorkorrutise abil defineeritud mitmemõõtmeliste momentide maatriksist erinevate elementide eraldamiseks.

Acta et Commentationes Universitatis Tartuensis, 968 ,1994, pp. 29-40.

Use of Absorption in Environmental Studies

T.Möls, A.-M.Parring

T.Möls, A.-M.Parring

Summary. A practical method based on absorption technique is proposed for analysis of environmental trends in the presence of a nuisance factor with a great number of levels. Addressed to ecologists, the paper presents both an illustrative example of discovering changes in Estonian small lakes and the mathematical basis of absorption technique

Key words: *absorption, ANOVA, environmental data analysis, nuisance factor, linear model, chemical oxygen demand, permanganate consumption*

1. Introduction and problem setting

This paper concerns environmental data analysis in cases where some categorical nuisance factors have too many levels to be handled with standard packages. We have a good opportunity to discuss these problems on the example of statistical analysis of the Data Bank of Estonian Small Waterbodies (DBESW) compiled by the Institute of Zoology and Botany in Tartu. The DBESW includes results of hydrochemical and physical analyses of about 4400 water samples from more than 350 Estonian lakes and other waterbodies investigated from 1925 to 1993. Collected partly during turbulent times in Estonia, the DBESW data are fragmentary and heterogeneous, different lakes and parameters being inspected at different non-overlapping periods. As a result, the corresponding data table is unbalanced and contains a high percentage of missing values. No good multivariate time series can be extracted from these data.

Estonian hydrochemists attempt currently to identify dominant changes in the water composition of local lakes. The problem is difficult because there exist significant individual variations between the lakes studied in different periods. Some tendencies observed in the data may be purely caused by these variations as well as by seasonal differences. At the same time, real changes may be completely camouflaged by distortions resulting from the bad (missing) experimental design.

It seems that the elimination of the effects of lakes and seasonality is an obligatory step in comparing different time periods. Roughly speaking, all measurements must be standardized by relating them to a fixed abstract standard lake and to some specific day within the year. The problem is how this elimination could be carried out in a situation where the total number of factorial effects exceeds 300 as in case of DBESW. Simple statistical packages fail to handle models with as many parameters.

One way to overcome the difficulty is to reduce the number of the levels of some factors. For example, in the trophic monitoring of Lake Peipsi (Möls et al, 1992) we have used clustering of control sites for reducing their number.

In this paper we propose another way to solve the problem within the analysis of variance. A key point in this approach is using the absorption technique for the elimination of the disturbing factor and taking into account only differences between the expected response values at different years or months. First, we give an example demonstrating absorption methods in environmental research. Futher, since our paper is meant as a reference material for hydrochemists not familiar with absorption and related topics, we explain these subjects more thoroughly in the next sections.

The present study was partially supported by the Estonian Science Foundation, Grant 73 (1993).

2. Model of organic components

In DBESW, 1118 records can be found with simultaneous measurements of chemical oxygen demand (CODCr) which characterizes the concentration of all organic components in water, and permanganate consumption (CODMn) as an indicator of mainly light organic substances. Every record in this selected data subset is characterized by the name of the lake (245 different names), year and the number of day within year (DN, ranging from 1 to 365).

Our goal is to investigate the dependence of $p = ln(\text{CODMn})$ on $ln(\text{CODCr})$, year and season. For this purpose we try to fit a

polynomial regression function to the data so that it would smooth out behavioral variations of lakes.

To enhance the analysis, the mean CODMn level of each studied lake must be taken into account. This can be achieved by eliminating the 'Lake' as a nuisance parameter from the model. Because the standard covariance analysis fails here as the number of levels of the factor 'Lake' is too large (245), absorption (explained in the next Section) will be made use of. After absorbing the factor 'Lake' by the SAS GLM procedure we can get the regression function represented in Table 1 (Model 1).

Table 1

Estimated coefficients and their standard errors (SE) in the regression function of p=ln(CODMn) after absorbing the nuisance factor 'Lake' (Model 1).

Here r = ln(CODCr), a = (year - 1925)/10, d = (number of day within year)/100.

Parameter	coefficient ± SE	Parameter	coefficient ± SE
d	-6.17559 ± 1.30060	r	-4.67018 ± 0.91880
d^2	-1.88400 ± 0.26879	r^2	1.21276 ± 0.27180
d^4	0.04430 ± 0.01473	r^3	-0.10625 ± 0.02783
a	-1.80148 ± 0.26364	a^2r	0.13891 ± 0.02273
a^4	0.00160 ± 0.00041	dar	-0.47143 ± 0.06352
da	7.47320 ± 1.26622	d^2r	0.73511 ± 0.11397
da^2	-2.30474 ± 0.44772	d^2ar	0.04904 ± 0.01189
da^3	0.33793 ± 0.06852	d^3r	-0.10358 ± 0.02400
da^4	-0.01730 ± 0.00379	a^3r	-0.01377 ± 0.00262
d^3a	0.04960 ± 0.01871	d^3r^2	-0.04144 ± 0.01069
d^3a^2	-0.00890 ± 0.00206	dar^2	0.02795 ± 0.00623

All coefficients in Model 1 are highly significant at the level $\alpha < 0.0001$ except coefficients for d^4 and d^3a with $\alpha = 0.0027$ and $\alpha = 0.0082$, correspondingly. General results of the analysis of variance of Model 1 are given in Table 2.

Table 2

Analysis of Variance Table

	DF	Sum of Squares	Mean Square	F
Model	268	387.87	1.447	42.15
Error	849	29.15	0.034	
Total	1117	417.02		

$R^2 = 0.93$, s = 0.185

It is important to stress that the meaning of the regression model estimated by means of absorption is not exactly to the same that naturalists mean when applying regression analysis. For example, the value of p predicted from Model 1 is not a real level of ln(CODMn) but only conditional. Particularly, the estimated model does not contain the intercept. Therefore, we shall refer to values calculated from Model 1 as pseudovalues. Related mathematical problems are discussed below in Section 5.

From Model 1 (see Table 1) various submodels can be derived. For example, by substituting $d = 2$ we can standardize the p value to a level which corresponds to the level of 19th July. For this certain day we get, from Model 1, the following Model 2:

$$p = -0.033a^4 - a^3(0.01378r - 0.67588) + a^2(0.1389r - 4.6808) + a(0.05592r^2 - 0.74669r + 13.542) - 0.106251r^3 + 1.047r^2 - 2.5584r.$$

This model is illustrated in Fig. 1 where the z-axis scale corresponds to pseudovalues. It can be used effectively only for calculating differences, not for getting perfect p values. However, a standard error of a pseudovalue p corresponding to a given year a, date d, and the CODCr level r can be calculated if the regression function of Model 2 is considered as a parametric function with coefficients a^4, a^3r etc. Using this method, we can calculate, with the help of SAS package, the standard error s_p for each predicted pseudovalue p. For example, from the Model 1 we get, for 19th July (corresponding to $d = 2$) and conditioned to the mean ln(CODCr) level $r = 3.36$, the results presented in Table 3.

Table 3

Some predicted pseudovalues p and their standard errors s_p. C denotes the undetermined constant.

Year	a	d	r	$C + p \pm s_p$
1950	2.5	2.0	3.36	-8.614±1.384
1960	3.5	2.0	3.36	-8.742±1.384
1970	4.5	2.0	3.36	-8.995±1.385
1980	5.5	2.0	3.36	-8.766±1.379
1990	6.5	2.0	3.36	-8.238±1.378
2000	7.5	2.0	3.36	-8.387±1.453

As one can see, standard errors s_p of the predicted pseudovalues are relatively large.

Let us now estimate changes in the p value during decades. Using the 'estimate' option in the SAS GLM procedure, Table 4 was drawn. The errors of differences are clearly much smaller than the errors of pseudovalues in Table 3. It seems that after absorption only differences have a practical value.

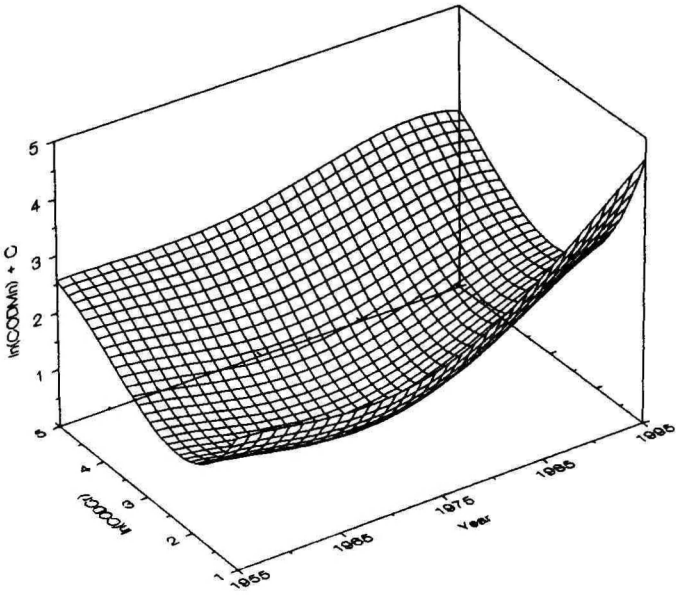

Figure 1. Graph of $p = ln(\text{CODn})$ as a function of $ln(\text{CODr})$, Year and Day. The Z-axis scale depends on undetermined constant C.

Table 4

Changes in p = ln(CODMn) level on 19th July as conditioned to ln(CODCr)=3.36, and the corresponding standard errors SE.

Time interval	Change in $p \pm SE$	Sign. level
1950 - 1960	-0.128 ± 0.043	0.0032
1950 - 1970	-0.381 ± 0.054	0.0001
1950 - 1980	-0.152 ± 0.055	0.0059
1950 - 1990	0.376 ± 0.040	0.0001
1950 - 2000	0.227 ± 0.317	0.4746
1960 - 1970	-0.254 ± 0.041	0.0001
1960 - 1980	-0.024 ± 0.054	0.6545
1960 - 1990	0.504 ± 0.039	0.0001

1970 - 1980	0.229 ± 0.042	0.0001
1980 - 1990	0.528 ± 0.042	0.0001
1990 - 2000	-0.149 ± 0.327	0.6840

3. The linear model and absorption

Let us have a linear model with one nuisance factorial variable with a huge number of levels and with some other arguments. If main interest is focused on parameters not connected with this factorial variable, it is appropriate to use the technique of absorption, see Searle S.R. (1971).

In order to understand what happens when absorption is used, let us consider a linear model with one factorial variable of k levels, which will be absorbed, and with p other arguments for which the parameters will be calculated. So let us consider the model

$$E\mathbf{Y} = \mathbf{F}\alpha + \mathbf{X}\theta,$$

where

- $\mathbf{Y}$ - $n \times 1$ vector of a dependent variable,
- $\mathbf{F}$ - $n \times k$ matrix of ones and zeros, associated with the factorial variable, to be absorbed,
- $\mathbf{X}$ - $n \times p$ matrix of arguments (factors or cofactors),
- α - $k \times 1$ vector of nuisance parameters,
- θ - $p \times 1$ vector of the parameters under study.

Let us denote the number of measurements on the level i of the nuisance factor by n_i and the number of all measurments by n, $n = \sum_{i=1}^{k} n_i$. The vector of the measured values of the dependent variable is denoted by $\mathbf{y}$.

The LSQ-estimate of the parameter-vector $\mathbf{g}$, $\mathbf{g} = [\alpha \vdots \theta]'$ is the solution of the normal equations

$$\mathbf{G}'\mathbf{G}\mathbf{g} = \mathbf{G}'\mathbf{y} \tag{1}$$

where

$$\mathbf{G} = [\mathbf{F} \vdots \mathbf{X}]. \tag{2}$$

It is clear that

$$\mathbf{G}'\mathbf{G} = \begin{bmatrix} \mathbf{F}'\mathbf{F} \vdots \mathbf{F}'\mathbf{X} \\ \cdots\cdots\cdots \\ \mathbf{X}'\mathbf{F} \vdots \mathbf{X}'\mathbf{X} \end{bmatrix}$$

where $\mathbf{F}'\mathbf{F}$ is a diagonal matrix with nonzero elements on the diagonal. These elements are numbers of measurements on the corresponding level of the nuisance variable,

$$\mathbf{F}'\mathbf{F} = Diag(n_1, n_2, ..., n_k). \tag{3}$$

From the first k equations of the linear system (1) it is possible to express the vector α. Indeed, as

$$\mathbf{F}'\mathbf{F}\alpha + \mathbf{F}'\mathbf{X}\theta = \mathbf{F}'\mathbf{y}$$

we get

$$\alpha = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'(\mathbf{y} - \mathbf{X}\theta). \tag{4}$$

Substituting the vector α into the last p equations of the system (1) yields

$$\mathbf{X}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{X}\theta = \mathbf{X}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{y}.$$

Let us calculate the $n \times n$ matrix $\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$. From (3) we get

$$(\mathbf{F}'\mathbf{F})^{-1} = Diag(\frac{1}{n_1}, \frac{1}{n_2}, ..., \frac{1}{n_k})$$

and

$$\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}' = \begin{bmatrix} \frac{1}{n_1}\mathbf{1}_{n_1 \times n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{1}_{n_2 \times n_2} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \frac{1}{n_k}\mathbf{1}_{n_k \times n_k} \end{bmatrix}$$

where $\mathbf{1}_{n_i \times n_i}$ is the $n_i \times n_i$ matrix of ones. Hence the difference $\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$ is a block-diagonal matrix

$$\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}' = \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{H}_k \end{bmatrix}$$

where $\mathbf{H}_i = \mathbf{I}_{n_i \times n_i} - \frac{1}{n_i}\mathbf{1}_{n_i \times n_i}$. The latter matrix is usually called the centering matrix, see Mardia, Kent, Bibby (1979); it is a projection matrix, i.e. it is symmetric and idempotent, $\mathbf{H}_i\mathbf{H}_i = \mathbf{H}_i$.

The matrix $\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$ may be named the *level-centering matrix.*

Let the matrix $\mathbf{X}$ be divided into blocks $\mathbf{X_i}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{bmatrix},$$

where the $n_i \times p$ block $\mathbf{X_i}$ consists of the values of arguments measured on the i-th level of the nuisance factor. Then the matrix

$$\dot{X} = \mathbf{H_i X_i} \tag{5}$$

is the $n_i \times p$ matrix of the centered values of arguments on the i-th level of the nuisance factor.

Let us denote by $\dot{\mathbf{X}}$ the matrix composed of blocks $\dot{\mathbf{X}}_\mathbf{i}$,

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{X}}_1 \\ \vdots \\ \dot{\mathbf{X}}_k \end{bmatrix}.$$

As

$$\mathbf{X_i'H_iX_i} = \mathbf{X_i'H_i'H_iX_i},$$

we obtain the equation

$$\mathbf{X}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{X} = \dot{\mathbf{X}}'\dot{\mathbf{X}}.$$

In the same way, we can divide the vector $\mathbf{y}$ into blocks

$$\mathbf{y} = \begin{bmatrix} \mathbf{y_1} \\ \vdots \\ \mathbf{y_k} \end{bmatrix}$$

where $\mathbf{y_i}$ is the vector of values of the dependent variable corresponding to the i-th level of the nuisance factor.

Similarly to the above procedure, the vector $\mathbf{H_i y_i}$ is the $n_i \times 1$ vector of the centered values of the dependent variable on the i-th level of the nuisance factor. Let us denote this vector by $\dot{\mathbf{y}}_\mathbf{i}$ and

$$\dot{\mathbf{y}} = \begin{bmatrix} \dot{\mathbf{y}}_1 \\ \vdots \\ \dot{\mathbf{y}}_k \end{bmatrix}.$$

Now we get

$$\mathbf{X}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{y} = \dot{\mathbf{X}}'\dot{\mathbf{y}}$$

and the linear system for the LSQ-estimate of θ is the following

$$\dot{\mathbf{X}}'\dot{\mathbf{X}}\theta = \dot{\mathbf{X}}\dot{\mathbf{y}}.$$

The complexity of this system does not depend on the number of levels of the absorbed nuisance factor; the system consists only of p unknown parameters.

Now we have got the following useful result.

Conclusion 1. *Let us have a linear model with one nuisance factor of k levels and p other arguments. If we absorb the nuisance factor, the LSQ-estimate for the vector θ of coefficients of other arguments will be given as a solution of a $p \times p$ system of linear equations*

$$\dot{\mathbf{X}}'\dot{\mathbf{X}}\theta = \dot{\mathbf{X}}\dot{\mathbf{y}}$$

where the dot marks level-centered data. Hence, we get

$$\hat{\theta} = (\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-}\dot{\mathbf{X}}\dot{\mathbf{y}}.$$

4. The characteristics of quality

Without any characteristics of the model quality the estimates of a linear model are usually useless. The most common characteristcs are the multiple coefficient of determination and the standard error. It is possible to calculate these characteristics using only the hitherto calculated matrices.

At first, let us calculate the sum of squared errors. If we suppose that normal equations have a solution $\hat{\mathbf{g}}$, we get

$$SSE = (\mathbf{y} - \mathbf{G}\hat{\mathbf{g}})'(\mathbf{y} - \mathbf{G}\hat{\mathbf{g}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{G}\hat{\mathbf{g}}.$$

Substituting $\mathbf{G}$ from (2) and $\hat{\mathbf{g}}$ from (4) we have

$$\begin{aligned} SSE &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{G}\mathbf{g} = \mathbf{y}'\mathbf{y} - \mathbf{y}'(\mathbf{F}\alpha + \mathbf{X}\theta) \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{X}\theta \\ &= \dot{\mathbf{y}}'\dot{\mathbf{y}} - \dot{\mathbf{y}}'\dot{\mathbf{X}}\hat{\theta} \end{aligned}$$

Thus, we get the second useful result.

Hence, the standard error is the following

$$s = \sqrt{\frac{SSE}{n - rank(G)}}$$

and the multiple coefficient of determination equals to

$$R^2 = 1 - \frac{SSE}{\dot{\mathbf{y}}'\dot{\mathbf{y}} + w}$$

where, using the common notation, $w = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$.

Also, it is possible to fill the ANOVA table for the whole model, which contains:

the sum of squares of the **model**	$\dot{\mathbf{y}}'\dot{\mathbf{y}} + \mathbf{w} - \mathbf{SSE}$
the sum of squared **errors**	SSE
the **total** sum of squares	$\dot{\mathbf{y}}'\dot{\mathbf{y}} + w$

5. Prognosis after absorption

It is clear that after the absorption of some parameters it is impossible to calculate residuals and the diagnostic statistics related to them. Let us examine possibilities to calculate the covariance matrix for the estimated parameter vector θ. It is well known that if the solution of normal equations (1) is unique, the estimation of the covariance matrix of $\hat{\mathbf{g}}$ is the following

$$D\hat{\mathbf{g}} = s^2(\mathbf{G}'\mathbf{G})^{-1}.$$

Using formulas for inverting a partitioned matrix (see Appendix) we get

$$\begin{aligned} D\hat{\theta} &= s(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{X})^{-1} \\ &= s(\mathbf{X}'(\mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}')\mathbf{X})^{-1} = s(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}. \end{aligned}$$

It is possible to calculate from a given change of other arguments prognosis and the standard error for a change of the dependent variable for a fixed level of the nuisance factor. Let us fix two values of the argument vector, $(x_1^{(1)}, \ldots, x_p^{(1)})'$ and $(x_1^{(2)}, \ldots, x_p^{(2)})'$. Then the change in arguments is vector $\mathbf{a}$, assuming the nuisance factor is fixed, we get

$$\mathbf{a} = (x_1^{(1)} - x_1^{(2)}, \ldots, x_1^{(1)} - x_p^{(2)})'.$$

values of the argument vector, $(x_1^{(1)}, \ldots, x_p^{(1)})'$ and $(x_1^{(2)}, \ldots, x_p^{(2)})'$. Then the change in arguments is vector **a**, assuming the nuisance factor is fixed, we get

$$\mathbf{a} = (x_1^{(1)} - x_1^{(2)}, \ldots, x_1^{(1)} - x_p^{(2)})'.$$

Now the prognosis for a change of the dependent variable is

$$\hat{\Delta}_y = \mathbf{a}'\theta$$

and the standard error for this prognosis is

$$s_{\hat{\Delta}} = s\sqrt{\mathbf{a}(\dot{\mathbf{X}}'\dot{\mathbf{X}})^{-1}\mathbf{a}}.$$

Acknowledgement. The authors would like to thank Dr. Henno Starast and Dr. Aini Lindpere for useful consultations and for permission to use data from DBESW.

Appendix

Suppose matrix **A** is partitioned,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

If all the necessary inverses exist, then the elements of $\mathbf{A}^{-1}$ will be

$$\mathbf{A}^{11} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}, \qquad \mathbf{A}^{12} = -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1},$$

$$\mathbf{A}^{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}, \qquad \mathbf{A}^{22} = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}.$$

References

Draper N.R., Smith H. (1981) Applied Regression Analysis, John Wiley & Sons

Mardia K.V., Kent J.T., Bibby J.M. (1979) Multivariate Analysis, Academic Press

Möls T., Saan T., Starast H., Lindpere A. (1992) Trophic monitoring of Lake Peipsi, Tallinn (in Estonian)

SAS/STAT User's Guide. Release 6.03 Edition, 1988, by SAS Institute Inc., Cary, NC, USA

Searle S.R. (1971) Linear Models, John Wiley & Sons

Received February, 1994

Absorbeerimise kasutamine keskonnauuringus

T.Möls, A.-M.Parring

Kokkuvõte

Artiklis tutvustatakse praktilist võtet – absorbeerimist – lineaarse mudeli parameetrite hindamiseks juhul, kui mudelisse kuulub suure tasemete arvuga segav faktor. Niisugune mudel osutus otstarbekaks kirjeldamaks muutuseid Eesti järvede seisundis ajavahemikul 1925-1993 kogutud andmete põhjal. Antakse lühiülevaade mudelist ja avatakse selle tehnika matemaatiline külg.

Acta et Commentationes Universitatis Tartuensis, **968** ,1994, pp. 41-54.

Correspondence Analysis as a Method for Depicting Qualitative Data

Kalev Pärna, Anneli Kulu

Summary. Correspondence analysis is an exploratory method for analysing 2-dimensional data tables where the rows and columns of the table are depicted as points in a low-dimensional (usually 2-dimensional) vector space. This paper gives an introduction to correspondence analysis covering essentials of the mathematical theory and demonstrating how the method can be applied to analyse a real data set.

Key words: *correspondence analysis, dual scaling, reciprocal averaging, singular value decomposition, contingency tables.*

1. Introduction

Nowadays statistics is used in very different fields and often a researcher tackles non-numerical variables. Actually, in social sciences, market research we deal almost exclusively with categorical data. Some variables of this type are: profession, country, type of car, region, sex, religious affiliation, race, agreement-disagreement with a given statement etc. Categorical data is usually represented in the form of contingency tables.

Correspondence analysis is a technique that in its simplest form can be applied to a two-way contingency table resulting in numerical values for both row and column categories. These values (scores) are chosen in such a way as to account for as much of the association between the two variables as possible. Then the row and column categories are displayed in 2-dimensional plots defined by pairs of these scores. It should be mentioned that the method can be applied not only for two-way contingency tables but to analyse large variety of data that can be brought into the form

of two-way table of non-negative numbers.

The method of correspondence analysis is very popular in France, particularly due to J.-P. Benzécri and his school. They also introduced the term *l'analyse des correspondances* (see Benzécri (1973)). But the history of the idea is longer and goes back as far as Hirschfeld (1935) and Fisher (1940). Their aim was to replace the two categorical variables representing certain biometrical data by numerical variables ('scores') in such a way as to maximize correlation between them. At the same period, but in a totally different (psychometric) context, L.Guttman (1941) derived a method for quantification of multivariate nominal data that in the case of two variables lead mathematically to the same solution as Fisher's method. Various names as 'canonical analysis of contingency tables', 'reciprocal averaging', 'scalogram analysis', 'dual scaling', and some other have been used to designate essentially the same method. Correspondence analysis (in its 'French' sense) can be regarded as a geometric form of the above approaches.

Correspondence analysis is mainly used as an exploratory method, and not too much attention has been paid (especially, in applications) to its statistical properties. This attitude is justified in many situations where the data can not be regarded as a random sample from any parent population. Researchers are more interested in stability properties of the method: do the small changes in data cause only small changes in the results, or not?

We suggest Greenacre (1984) for a detailed exposition of the method and Greenacre, Hastie (1987) for an overview.

2. The method of correspondence analysis

Here we give a description of the mathematical method that stands behind the correspondence analysis (see also the section 4).

Let us have n observations (persons, objects) classified by two categorical variates A and B according to their categories $A_1, A_2, \ldots, A_I$ and $B_1, B_2, \ldots, B_J$. The number of observations having simultaneously both values A_i and B_j will be denoted by n_{ij}. The data can be presented in the form of IxJ contingency table $N = (n_{ij})$ which row, column, and grand totals are $n_{i.} = \Sigma_j n_{ij}, n_{.j} = \Sigma_i n_{ij}$, and $n = \Sigma_i \Sigma_j n_{ij}$, respectively. Let us denote two diagonal matrices $R = diag(n_{1.}, \ldots, n_{I.}), C = diag(n_{.1}, \ldots, n_{.J})$ and relative frequencies $f_{ij} = n_{ij}/n$, $f_i = n_{i.}/n$, $f_j = n_{.j}/n$.

In correspondence analysis our main interest is with conditional laws in the contingency table N. Let us denote row-wise and

column-wise relative frequencies (conditional probabilities) by:

$$f_j^i = n_{ij}/n_{i.}, \qquad f_i^j = n_{ij}/n_{.j}.$$

The vector $f_B^i = (f_1^i, f_2^i, \ldots, f_J^i)^T$ will be called the *row profile* (for the row i). Similarly, the vector $f_A^j = (f_1^j, f_2^j, \ldots, f_I^j)^T$ will be called the *column profile* (for the column j). The looser term 'profile' (instead of 'conditional law') has the advantage that it can be applied to any rectangular table N of non-negative numbers (not necessarily contingency tables). Observe that row profiles f_B^i represent I points in the J-dimensional Euclidean space. We will call these points a *cloud of row profiles* and denote it $N_B(A)$:

$$N_B(A) = \{f_B^i \mid i = 1, \ldots, I\}.$$

It is essential to correspondence analysis that each point f_B^i in the cloud has its *mass* defined by the marginal probability f_i. So, the cloud is a configuration of weighted points. The *centroid* of the cloud is defined as its mass-center, i.e. the weighted average of its elements:

$$f_B = \Sigma_i f_i \cdot f_B^i$$

which is readily seen to be equal to the marginal distribution of the variable B:

$$f_B = (f_1, \ldots, f_J)^T.$$

Let us make a comment on the dimensionality problem. Initially, all points of the cloud $N_B(A)$ are the elements of J-dimensional space. But since the coordinates of these vectors sum up to unity (conditional laws!) we see that the profiles actually lie on a hyperplane of dimension $J-1$. From the other side , any I points can be amounted by a subspace of dimension $I-1$. Thereby, the cloud needs no more than $\min\{J-1, I-1\}$ dimensions to be amounted. Actual dimensionality can be even lower, depending on the initial data. For example, we loose one more dimension if the data table N contains two proportional rows. The 'right' dimensionality K of the cloud is determined by the rank of the data matrix N : $K =$ rank(N)$-1 \leq \min\{I-1, J-1\}$.

Now let us consider the *cloud of column profiles*:

$$N_A(B) = \{f_A^j \mid j = 1, \ldots, J\},$$

a subset of an I-dimensional space. The elements of $N_A(B)$ have masses equal to corresponding marginal probabilities f_j, and the centroid of the cloud is the weighted average of its elements which is now equal to the marginal distribution of the variable A:

$$f_A = (f_1, \ldots, f_I)^T.$$

We can now formulate the aim of correspondence analysis. In geometrical terms, our aim is to *identify a low-dimensional subspace which comes 'closest' to all points of the cloud* (Greenacre (1984)). The general idea is the same as in principal component analysis but also several differences can be noticed after having specified the term 'closeness'. If the points have different masses f_i then the subspace should lie even closer to the points of higher mass, while a deviation from the points of lower mass would be more easily tolerated. Also, the distance we use is not common but a weighted Euclidean distance (see below).

As a simple illustration let us have a data in the form of (8x3)-contingency table. Since the eight row profiles lie on a 2-dimensional simplex defined by

$$f_1^i + f_2^i + f_3^i = 1, \text{ all } f_j^i \geq 0,$$

the cloud of 8 row profiles can be depicted as on Fig.1.

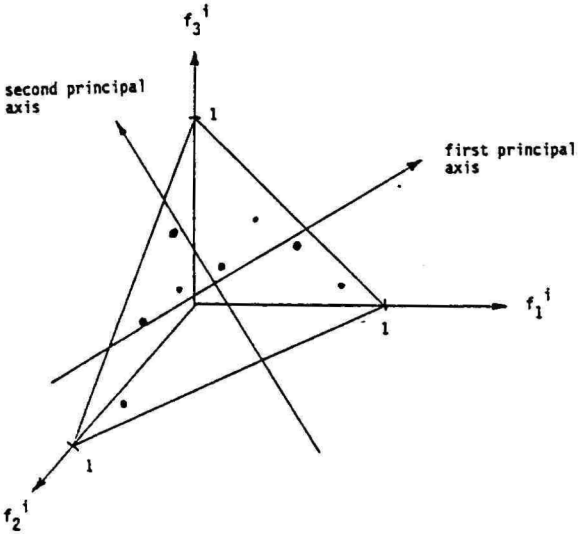

Fig.1. All eight points (the row profiles) lie on the triangle.

We now specify two important notions. The *distance* of the row profile f_B^i from the centroid of the cloud is, by definition,

$$d^2(i) = \Sigma_j (f_j^i - f_j)^2 / f_j, \tag{2.1}$$

which is called the 'chi-squared distance'. In fact, it is a squared *weighted Euclidean distance*, since the normalizing factors $\sqrt{f_j}$ are used to measure the difference between the j-th co-ordinates. The same in matrix form is:

$$d^2(i) = (f_B^i - f_B)^T n C^{-1} (f_B^i - f_B). \tag{2.1'}$$

The crucial notion in correspondence analysis is inertia, a generalization of variance. ***Inertia of the cloud*** $N_B(A)$ is defined by the weighted average of the distances of$I$ profiles from the centroid of the cloud:

$$in(A) = \Sigma_i f_i \cdot d^2(i). \tag{2.2}$$

Inertia is the measure of how much the profiles are spread around the centroid. In the special case when all masses f_i are equal, inertia reduces to the total variance of the$I$ points. Inertia is closely related to chi-squared statistics for testing the independence in the two-way contingency table. Namely, it is elementary to show that

$$in(A) = \Sigma_i \Sigma_j \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}}$$

$$= \chi^2 / n. \tag{2.3}$$

As to the dual cloud $N_A(B)$, its inertia is determined by

$$\text{in}(B) = \Sigma_j f_j \cdot d^2(j)$$

where

$$d^2(j) = \Sigma_i (f_i^j - f_i)^2 / f_i$$

- the weighted Euclidean distance of the column profiles f^j from the centroid f_A . The formula (2.3) also applies to the dual cloud, and we denote the common value of the two inertias by λ:

$$\lambda = in(A) = in(B) = \chi^2 / n. \tag{2.3'}$$

Our gain in correspondence analysis - a specific low-dimensional sub-space which comes close to the points of the cloud - is determined by *principal axes of inertia*. Principal axes are K vectors

applied to the centroid of the cloud and showing the directions of largest inertia, every next being orthogonal to all previous. To get a low-dimensional space only a small number (say, $K^* = 2$) of first principal axes are needed, in hope that these K^* dimensions describe the data well enough (see Fig.1). The complete mathematical solution to the problem is embodied in the concepts of *singular value decomposition* and *low rank approximation* (see Section 4 for details). Here we only give some facts that are important from applied point of view.

Let $m_1, \ldots, m_K$ be the principal axes of inertia for the cloud of the row profiles and let x_{ik} be the co-ordinate of the i-th row profile with respect to the k-th principal axis of inertia $-$*k-th principal co-ordinate*. Then the i-th row profile is given by the K-vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ and the whole cloud of row profiles by an $(I\text{x}K)$-matrix X,

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_I \end{pmatrix}$$

It comes mathematically from the singular value decomposition method that the K columns of the matrix X can be computed as eigenvectors of the $(I\text{x}I)$-matrix $R^{-1}NC^{-1}N^T$ corresponding to its non-zero eigenvalues $\lambda_1, \ldots, \lambda_K$. It means that the matrix X satisfies the equation

$$(R^{-1}NC^{-1}N^T)X = XD_\lambda \tag{2.4}$$

where D_λ is a $(K\text{x}K)$-diagonal matrix of eigenvalues, D_λ = diag $(\lambda_1, \ldots, \lambda_K)$, assuming the ordering $\lambda_1 \geq, \ldots, \geq \lambda_K > 0$. The equation (2.4) is equivalent to (4.9), but because of symmetricity the latter is computationally more convenient to solve. Observe also that, actually, the number of non-zero eigenvalues in (2.4) equals rank$(N) = K + 1$. However, we do not use the 'trivial' largest eigenvalue $\lambda_0 = 1$ and the corresponding constant eigenvector consisting of $1's$, which is easily seen to satisfy (2.4) but not of real interest. Geometrically, the dropping of the trivial eigenvalue and eigenvector means that in our analysis we set the origin of principal axes at the centroid of the cloud.

If we normalize the eigenvectors so that

$$\frac{1}{n}X^T RX = D_\lambda \tag{2.5}$$

then the weighted sum of squares along the k-th principal axis is equal to the eigenvalue λ_k - called, thereby, the k-th *principal inertia.*

In the case of the dual cloud let the principal co-ordinates of the j-th column profile be $y_{j1}, y_{j2}, \ldots, y_{jK}$ which are collected into the $(J \mathrm{x} K)$-array Y. However, it is not necessary to solve a new eigenvalue problem (an analog of (2.4) - instead we can exploit simple 'transition formulae' between the two sets of principal co-ordinates:

$$YD_\rho = C^{-1}N^T X, \tag{2.10}$$

$$XD_\rho = R^{-1}NY \tag{2.11}$$

where $D_\rho = D_\lambda^{1/2}$. The use of (2.10) results in y-co-ordinates that have similar to x-co-ordinates standardization (see the formula (4.11). The resulting equality of principal inertias in both clouds allows to merge the two separate graphical displays into one joint display with I points for row and J points for the column profiles. Still, according to Greenacre (1984) we should 'avoid the danger of interpreting distances between points of different clouds, since no such distances have been explicitly defined'.

Finally, we show that the principal inertias $\lambda_1, \lambda_2, \ldots, \lambda_K$ are, in fact, the components of the total inertia λ. Indeed, we have

$$n\lambda = \chi^2 = n(\Sigma_i \Sigma_j \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1)$$

$$= n(tr(R^{-1}NC^{-1}N^T) - 1)$$

$$= n(\lambda_1 + \lambda_2 + \ldots + \lambda_K), \tag{2.12}$$

since the trace of the matrix is equal to the sum of its eigenvalues. Therefore $\lambda = \mathrm{in(A)} = \mathrm{in(B)} = \lambda_1 + \lambda_2 + \ldots + \lambda_K$ - the total inertia is divided between K principal axes, the first axis taking the largest portion etc. In computer outputs the *relative inertias* λ_k/λ are usually given (expressed in percents) to show relative importances of different axes in describing the data.

3. An application of correspondence analysis

A sample of 105 university students were asked to characterize their teachers in order to get some idea about what is important in teachers from students' point of view. A complete list of 58

teachers of a faculty was presented to each student who had to find one or more characteristic features for every teacher in the list. In order to get an 'unbiased' data the question was formulated in a very loose form, no prior list of possible features was used. It was allowed to skip if the student did not know some of the teachers. The total number of different words used by the students exceeded 80, but in order to supress the dimensionality the words with almost the same meaning were replaced by one of them. Thereby, as many as 55 qualities (listed below) were used in the analysis.

All the data were collected into a 58x55 contingency table showing how many times one or other word (column of the table) was used to characterize a specific teacher (a row of the table). The data table is too large to be presented here.

List of characteristics used by students

A slow	a nervous
B active	b objective
C authoritarian	c optimistic
D authoritative	d ill-posed
E vain	e pedant
F egoistic	f gabbler
G emotional	g stubborn
H conceited	h quiet
I fanatic	i strict
J absent-minded	j glad
K well-wishing	k confusing
L sense of humour	l sarcastic
M dull	m solid
N talkative	n subjective
O closed	o communicative
P fast	p friendly
Q correct	q likable
R careless	r reserved
S kind	s balanced
T pleasant	t moody
U nonunderstanding	u serious
V enjoyable	v dizzy
W understanding	w trusting
X naive	x complaisant
Y censorious, nagging	y polite
Z demanding	z whiner
! overbearing	" undemanding

* supercilious

The correspondence analysis procedure of the MVSP statistical package was used to analyse the data. Because of the large number of points the emphasis is put on graphical display rather than numerical analysis. For the same reason two separate displays are given, although, in principle, they can be overlayed.

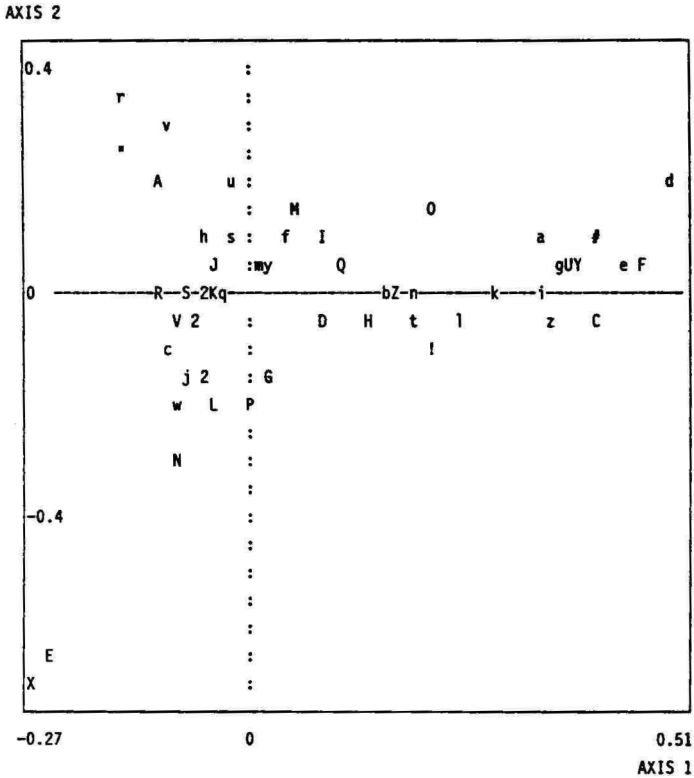

Fig.2. Display of the characteristics
(2 = two overlapping characteristics)

In this example the first and the second principal axes describe 14 and 11 percent, respectively, of the total inertia involved in the data (thus $\lambda_1/\lambda = 0.14, \lambda_2/\lambda = 0.11$). The numbers are

not very large, but it should be taken into consideration that the dimensionality of the data can be as large as 54 and, therefore, the total inertia is distributed between many (maximum 54) axes. In a hypothetical extreme case where the total variance were uniformly distributed between all 54 axes the first two axes would cover only 3.8 percent of the total variance, which is 7 times less than in our real case. Moreover, the two dimensions are easy to interprete.

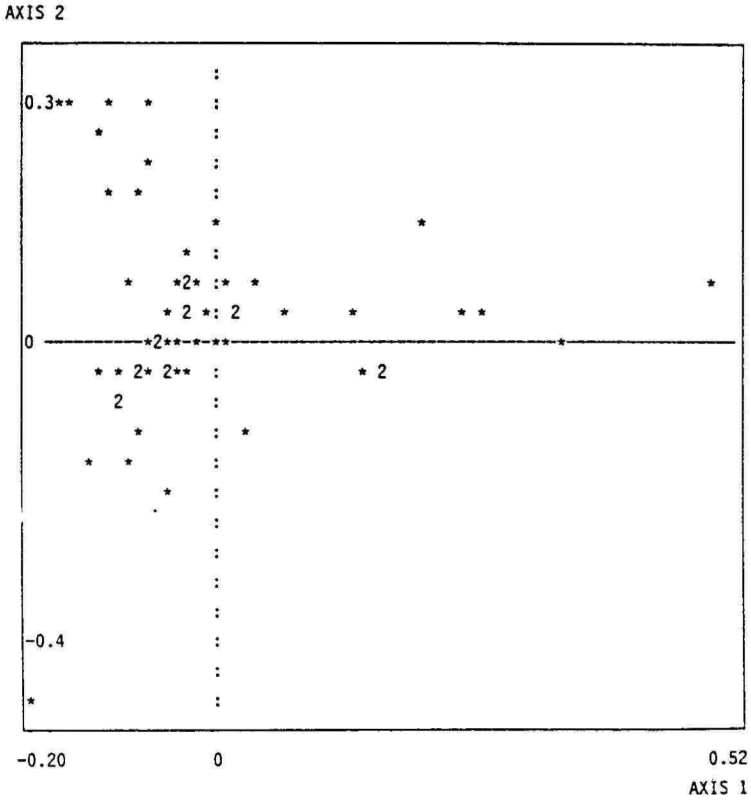

Fig.3. Display of the teachers
2 = two overlapping teachers

On the display of characteristics (see Fig.2.) one notices a long right-hand tail with various words expressing demandingness of teachers, whereas the words on the left-hand side have opposite

meaning. Since it is the first principal axis we conclude that, from students' point of view, the strictness is the most important feature that discriminates between university teachers.

According to the meaning of the characteristics at the top and the bottom of the same display, the second principal axis can probably be called the introversy-extraversy dimension ('closed' people - 'open' people).

Distribution of the points on the teachers' display has approximately the same shape (see Fig.3). Overlaying the two displays one can (to some extent) decide on which characteristics are related with any specific teacher.

4. Appendix: Singular value decomposition and low rank approximation

A fundamental theorem of matrix algebra states that each rectangular matrix permits a singular value decomposition (SVD), as explained below.

Let $\mathbf{A}$ be an $(I\text{x}J)$ data matrix, and $\mathbf{G}(I\text{x}I)$ and $\mathbf{H}(J\text{x}J)$ be given positive definite symmetric matrices. The *generalized SVD* of $\mathbf{A}$ is the following decomposition of $\mathbf{A}$:

$$\mathbf{A} = \sum_{k=1}^{K} \rho_k l_k m_k^T. \tag{4.1}$$

where $l_1, \ldots, l_K$ are orthonormalized (in metrics $\mathbf{G}$) I-vectors, $m_1, \ldots, m_K$ are orthonormalized (in metrics $\mathbf{H}$) J-vectors and $\rho_1, \ldots, \rho_K$ are positive numbers (in decreasing order). The vectors l_k and m_k are uniquely determined up to the reflections. In view of form (4.1) of the SVD it seems that if, for some K^*, singular values $\rho_{K^*+1}, \ldots, \rho_K$ are relatively small compared to $\rho_{K^*}, \ldots, \rho_K$, then dropping the last $K - K^*$ terms of the right hand side of (4.1) gives a good approximation to $\mathbf{A}$ and has lower rank than $\mathbf{A}$. This approximation

$$\mathbf{A}_{(K^*)} = \sum_{k=1}^{K^*} \rho_k l_k m_k^T \tag{4.2}$$

turns out to be a generalized least squares one, and this is the result which makes the SVD so useful. More precisely, $\mathbf{A}_{(K^*)}$ is the best approximator in the sense that it minimizes:

$$tr\{(\mathbf{G}(\mathbf{A} - Z)\mathbf{H}(\mathbf{A} - Z)^T)\}. \tag{4.3}$$

over all possible IxJ-matrices Z of rank K^*(see Kshirsagar (1972)). The same SVD in matrix form looks like

$$\mathbf{A} = LD_\rho M^T, \text{ with } L^T\mathbf{G}L = M^T\mathbf{H}M = \mathbf{I}. \tag{4.4}$$

where the columns of L and M are l_k and m_k respectively.

Let us apply the theory for the purposes of correspondence analysis now. As stated in the section 2, in correspondence analysis our aim is to find a low-dimensional subspace that comes closest to all points in a given cloud. Let us consider the cloud of the row profiles f_B^i. First, it can be shown that the 'best' subspace must contain the centroid of the cloud, i.e. the marginal profile vector f_B. (The best 0-dimensional subspace is the point f_B itself!) Therefore, instead of the row profiles we will approximate the centred row profiles $f_B^i - f_B$, i.e. the rows of the matrix

$$\mathbf{A} = R^{-1}N - 1_I f_B^T, \tag{4.5}$$

where 1_I is the column of I unities. Each row of $\mathbf{A}$ has its mass f_i, and the distances are defined by the diagonal matrix C(see the formula (2.1')). Thereby, the criterion for finding the best approximation for $\mathbf{A}$ is: minimize

$$\sum_i f_i(f_B^i - f_B - z_i)^T nC^{-1}(f_B^i - f_B - z_i) =$$

$$= tr\{(1/n)R(R^{-1}N - 1_I f_B^T - Z)nC^{-1}(R^{-1}N - 1_I f_B^T - Z)^T\} \tag{4.6}$$

over all IxJ-matrices Z of rank K^*. It means that the rows of Z are unknown points z_i that must belong to a K^*-dimensional subspace. This function is a particular case of the criterion (4.3) with $\mathbf{G}$ and $\mathbf{H}$ as diagonal matrices:

$$\mathbf{G} = \frac{1}{n}R = \text{ diag } \{f_i\},$$

$$\mathbf{H} = (\frac{1}{n}C)^{-1} = \text{ diag } \{1/f_j\}.$$

Therefore, the solution is embodied in the generalized SVD of $\mathbf{A}$:

$$R^{-1}N - 1_I f_B^T = LD_\rho M^T \tag{4.7}$$

with normalization

$$L^T(\frac{1}{n}R)L = M^T(\frac{1}{n}C)^{-1}M = \mathbf{I}. \tag{(4.8)}$$

The columns of M, the vectors $m_1, \ldots, m_K$ (say), define the principal axes, and the rows of LD_ρ define the co-ordinates with respect to these axes, i.e. principal co-ordinates. (Recall that in the section 2 the principal co-ordinate matrix was denoted by X - thus we have $X = LD_\rho$.) Due to (4.8) we have normalization

$$\frac{1}{n} X^T R X = D_\lambda \tag{4.9}$$

which prescribes that the weighted sum of squares of the points' co-ordinates along the k-th principal axis is equal to the eigenvalue λ_k.

How to compute the principal co-ordinates matrix $X = LD_\rho$? Premultiplying (4.7) by $\frac{1}{n}R$ gives another generalized SVD problem:

$$\frac{1}{n} N - f_A f_B^T = \frac{1}{n} \mathrm{RLD}_\rho M^T, \tag{4.10}$$

which, in turn, can be shown to be equivalent to a symmetric eigenvalue problem

$$(R^{-1/2}\mathbf{N}C^{-1/2})(R^{-1/2}\mathbf{N}C^{-1/2})^T(R^{1/2}X) = (R^{1/2}X)D_\lambda \tag{4.11}$$

where $D_\lambda = D_\rho^2$, and $\mathbf{N} = N - nf_A f_B^T$. Actually, one can solve (4.11) with N instead of $\mathbf{N}$. Since the rank of N is by 1 larger than the rank of $\mathbf{N}$ we then will obtain one more eigenvalue and eigenvector (additional column to X). But it can be readily checked that the additional column will be 'trivial' in the sense that it consists of $1's$ only, and thus can be neglected as uninteresting. The trivial solution simply points the centroid of the cloud as its best 0-dimensional approximation.

Having obtained principal co-ordinates for the rows, those for the columns can be calculated *via* the following 'transition formula':

$$Y = C^{-1}N^T X D_\rho^{-1}, \tag{4.12}$$

which yields in standardization

$$\frac{1}{n} Y^T C Y = D_\lambda. \tag{4.13}$$

In view of identical right hand sides in (4.9) and (4.13), both clouds have the same principal inertia along each axis. This makes the row and column co-ordinates comparable, justifying their joint display.

References.

Benzécri, J.-P. (1973) L'Analyse des Données. T.1: L'Analyse des Correspondances. Dunod, Paris.

Fisher, R.A. (1940) The precision of discriminant functions. Ann. Eugen., **10**, 422-429.

Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis. Academic Press, London.

Greenacre, M., Hastie, T. (1987) The geometric interpretation of correspondence analysis. Journ. Amer. Statist. Assoc., **82**, 437-447.

Guttman, L. (1941) The quantification of a class of attributes: a theory and method of scale construction. In "The Prediction of Personal Adjustment" . Social Science Research Council, New-York, pp.319-348.

Hirschfeld, H.O. (1935) A connection between correlation and contingency. Cambridge Philosophical Soc. Proc. (Math. Proc.), **31**, 520-524.

Kshirsagar, A.M. (1972) Multivariate Analysis. Marcel Dekker, New-York.

Received February, 1994

Korrespondentsanalüüs kui kvalitatiivsete andmete kujutamise meetod

Kalev Pärna, Anneli Kulu

Kokkuvõte

Korrespondentsanalüüs ehk "prantsuse faktoranalüüs" on andmetöötlusmeetod, mis rakendatuna 2-mõõtmelisele sagedustabelile omistab tabeli ridadele ja veergudele teatud mõttes optimaalsed arvväärtused (st. kvantifitseerib rea- ja veerutunnused), mida kasutades saab tabeli ridu ja veerge kujutada punktidena väikese-mõõtmelises ruumis, näiteks tasandil. Arvväärtused valitakse nii, et nad võtaks arvesse nii palju seost kahe tunnuse vahel kui võimalik. Artiklis antakse ülevaade korrespondentsanalüüsi meetodist, näidates selle sarnasust hästituntud peakomponentide meetodiga ning seostades teda maatriksi singulaarse lahutuse mõistega. Kirjeldatakse meetodi kasutamist ühe üliõpilasküsitluse andmete analüüsimisel.

Acta et Commentationes Universitatis Tartuensis, **968** ,1994,pp. 55-75.

Using Index-vectors and Partitions in Multivariate Analysis

Ene-Margit Tiit, Sirli Tammet

Summary. The paper considers some concepts, which are used in different problems of multivariate analysis – the *index-vector*, i.e. vector, having natural components and identifying the subvector or submatrix, and the *partition* of an index-vector, identifying the division of a set of variables into several subsets. In the first part of the paper some properties of lexicographical ordering of index-vectors are represented, which are useful for estimating higher moments and their functions and solving the problem of 'moments of moments', too. The second part of the paper presents some algorithms for generating partitions of a number. All results are illustrated with examples.

Key words: *correlation matrix, mixed moments, index-set, partition of set, partition of number, extremal distribution.*

1. Introduction

In multivariate statistics the concept of an *index-vector* is widely used for identifying the subsets of variables or the components of random vectors, for defining higher moments and cumulants etc. In this paper several problems connected with describing and ordering of multivariate statistical objects, identified by the index-vectors, are regarded.

In the first part (Sections 2 – 3) of the paper some identification problems of the higher moments of multivariate random vectors are solved using the ordering rule of index-vectors. The theoretical discussion is illustrated with the help of an example, where the index-vectors, corresponding to all fourth moments of a five-variate random vector are found (Table 1).

The results received in Section 3 are generalized in Section 4, where the rule for identifying the 'moments of moments' is given, (see Fisher, 1929 and Kotz & Johnson, 1981, 124 - 129). In Table 2, the double index-vectors, identifying the third moments of the fourth moments of a bivariate random vector, are represented.

In the second part (Section 5) of the paper the concept of the partition of number is used for building a classification of higher moments of random vectors. Also a formula for calculating the cardinalities of classes of different moments is derived in this part. The theoretical results are illustrated by an example, where the data from Table 1 are used.

The third part (Sections 6 and 7) of the paper is devoted to the construction of partitions of a given index-vector. This problem is essential in several procedures of building statistical models (e.g. the regression analysis), but also in the construction of extremal distributions and extremal correlation matrices, see Tiit (1992). In Section 6 the algorithm for constructing of binary partitions is given and illustrated in Examples 4 and 5. In Section 7 the algorithm is generalized for building arbitrary partitions. The algorithm is illustrated in Example 6.

2. Index-vector

Let us start with the definitions of the basic concepts.

DEFINITION 1. The vector

$$I = (i_1, \ldots, i_k) \tag{1}$$

is said to be an *index-vector*, if all its components are integers, $i_j \leq p$, p is a fixed parameter.

The set of all index-vectors (1) will be denoted by $\Upsilon(p, k)$. Let us assume that the parameter p and the dimension k are fixed. Let us regard the equivalence classes of index-vectors by permutations of indices.

DEFINITION 2. Let $I = (i_1, \ldots, i_k)$ and $J = (j_1, \ldots, j_k)$ be two index-vectors from the class $\Upsilon(p, k)$. If there exists a permutation P so, that the equality $I = PJ$ holds, then we say that the index-vectors I and J are *equivalent*.

As *a representati ve* of a class of equivalent index-vectors we will take the *lexicographically decreasingly ordered* index-vector

$$I' = (i'_1, \ldots, i'_k)$$

from this class, where

$$i'_j \leq i'_{j-1},\ j = 2, \ldots, k. \tag{2}$$

The index-vectors that belong to different classes are said to be *essentially different.* In the future (Sections 3 and 4) we will regard the set $\Psi(p,k)$ of equivalence classes (essentially different index-vectors), $\Psi(p,k) \subset \Upsilon(p,k)$.

3. Ordering of index-vectors

Let us define the *position* $f(\cdot)$ of a component i_f of an index-vector I as its successive number in the lexicographically ordered index-vector $(i_1, \ldots, i_k)$. Every index-vector $(i_1, \ldots, i_k)$ can be regarded as an k-figure natural number in the positional number system having the basis p.

Let us order all index-vectors of the set $\Psi(p,k)$ lexicographically; as a result we receive the sequence $\mathcal{I}(p,k)$ with terms I_h (here I_h is the h-th element of the sequence). The sequence $\mathcal{I}(p,k) = \{I_h\}$ does not include all possible k-figure p-numbers, but only these which satisfy the conditions (2).

Let us assign to every index-vector I_h its *rank* $h = h(I)$. Our task is to find the ordering rule ϕ, describing the correspondence

$$\phi :\ I_h \Rightarrow h(I),\ \phi^{-1} : h(I) \Rightarrow I_h. \tag{3}$$

In the following we will define several concepts, connected with the sequence $\mathcal{I}(p,k)$.

DEFINITION 3. The subsequence of the sequence $\mathcal{I}(p,k)$, consisting of k-variate index-vectors $(i_1, \ldots, i_k)$, satisfying the condition

$$i_1 \leq j$$

is said to be a *j-segment* of the sequence $\mathcal{I}(p,k)$ and will be denoted by $H(j,k)$, $j = 1, \ldots, p$.

Let $\kappa_H(j,k)$ denote the *cardinality* of the segment $H(j,k)$.

DEFINITION 4. The subsequence of the sequence $\mathcal{I}(p,k)$, consisting of k-variate index-vectors $(i_1, \ldots, i_k)$, satisfying the condition

$$i_1 = j$$

is said to be a *j-section* of the sequence $\mathcal{I}(p,k)$ and will be denoted by $G(j,k)$, $j = 1, \ldots, p$.

Let $\kappa_G(j,k)$ denote the cardinality of the section $G(j,k)$.

From the definitions it follows that between the cardinalities of segments and sections the following equation holds:

$$\kappa_H(j,k) = \sum_{f=1}^{j} \kappa_G(f,k). \tag{4}$$

To find the ordering rule ϕ we will start with the simpliest possible index-vectors I.

A. Let us construct the sequence $\mathcal{I}(p,k)$ in the case $k=1$. Then $I = i_1$ and we have the following equations for the segments

$$H(j,1) = \{1,\ldots,j\},\ \kappa_H(j,1) = j,$$

and the sections of the sequence $\mathcal{I}(p,1)$:

$$G(j,1) = \{j\},\ \kappa_G(j,1) = 1,\ j = 1,\ldots,p.$$

In the case $k=1$ the ordering rule ϕ is the following:

$$h(I) = i_1,\ i_1 = 1,\ldots,p. \tag{5}$$

The cardinality of the sequence $\mathcal{I}(p,1)$ is p.

B. Let us consider the case $k=2$. In this case the members of the sequence $\mathcal{I}(p,2)$ are the vectors (i_1,i_2). Let us regard a section $G(j,2)$. By the conditions (2) and the lexicographical ordering rule of the sequence $\mathcal{I}(p,2)$ it can be concluded that the index-vectors (i_1,i_2), belonging to the section $G(j,2)$, have the following construction:

$i_1 = j$,

i_2 has all successive values of the index i'_1, belonging to the segment $H(j,1)$ of the sequence $\mathcal{I}(p,1)$.

From this construction the expression of the cardinality of a section immediately follows:

$$\kappa_G(j,2) = \kappa_H(j,1),\ j = 1,\ldots,p. \tag{6}$$

From this construction another important statement follows, too. If the vector $I = (i_1,i_2)$ belongs to the section $G(i_1,2)$, then for the rank $h(I)$ the following inequalities:

$$\kappa_H(i_1-1,2) < h(I) \leq \kappa_H(i_1,2),\ i_1 = 1,\ldots,p. \tag{7}$$

are valid.

To get the precise value of the rank $h(I)$ we will regard all index-vectors (j, i_2) from the section $G(j, 2)$. It is easy to see that there exists *one-to-one correspondence* between the section $G(j, 2)$ and the sequence $\mathcal{I}(j, 1)$, formed from the second components $\{i_2\}$ of the initial index-vectors. Using the equation (5), we will get from the equation (7) the following formula:

$$h(I) = \kappa_H(i_1 - 1, 2) + i_2.$$

The last equation can be transformed with the help of the formulae (4) and (6):

$$\kappa_H(j, 2) = \sum_{f=1}^{j} \kappa_H(f, 1) = ((j \cdot (j + 1))/2 = A_2^j,$$

where we use the notation $A_g^f = ((f + g - 1) \cdot \ldots \cdot f)/(g!)$.

Using the equation $A_1^g = g$, it will be convenient to rewrite the formula of the transformation rule ϕ in the following, more symmetric (in the sense of indices), shape:

$$h(I) = 1 + A_2^{i_1 - 1} + A_1^{i_2 - 1}. \tag{8}$$

The cardinality $\kappa(\mathcal{I}(p, 2)$ can be calculated with the help of equation (8):

$$\kappa(\mathcal{I}(p, 2)) = 1 + A_2^{p-1} + A_1^{p-1} = A_2^p. \tag{9}$$

C. Let us prove now Theorem 1, determining the rule ϕ for the arbitrary dimensionality k.

THEOREM 1. Let $\Psi(p, k)$ be the set of essentially different index-vectors $(i_1, \ldots, i_k)$, where the dimensionality k and the parameter p are fixed. The rule ϕ for the lexicographical ordering of index-vectors in the sequence $\mathcal{I}(p, k)$ is the following:

$$h(I) = 1 + \sum_{g=1}^{k} A_{k+1-g}^{i_g - 1}. \tag{10}$$

PROOF. The formula (10) is an immediate generalization of the formula (8) by the principle of mathematical induction.

On the first step of the proof we fix the *position* of the index-vector $I = (i_1, \ldots, i_k)$ in the sequence $\mathcal{I}(p, k)$, using the concepts of segments and sections of the sequence.

By Definition 3 and 4 the index-vector $(i_1, \ldots, i_k)$ belongs to the section $G(i_1, k)$ and hence to the segment $H(i_1, k)$ of the $\mathcal{I}(p, k)$, as well. Let us take the section $G(i_1, k)$ and regard all $(k-1)$-variate subvectors of the index-vectors $(i_1, \ldots, i_k)$ belonging to the section $G(i_1, k)$ and having different values on positions $2, \ldots, k$. This sequence coincides with the sequence $\mathcal{I}(p, k-1)$. The $(k-1)$-subvector $(i_2, \cdots, i_k)$ of the initial index-vector I belongs to the segment $H(i_2, k-1)$ of the sequence $\mathcal{I}(p, k-1)$.

Repeating this discussion, we get on the f-th stage a sequence $\mathcal{I}(p, k+1-f)$ of $(k+1-f)$-variate subvectors of index-vectors from the set $\Psi(p, k)$, and fix the segment $H(i_f, k+1-f)$, where the $(k+1-f)$ subvector of the initial index-vector I belongs, $f = 1, \cdots, k$. From here we receive the formula for h(I):

$$h(I) = \sum_{f=1}^{k-1} \kappa_H(i_f - 1, k+1-f) + i_k. \tag{11}$$

The next step is to calculate the cardinalities $\kappa_H(i_1 - 1, k+1-f)$. From the equations (4) and (6) and the construction of the sequence $\mathcal{I}(p, k)$ the following equation can be concluded:

$$\kappa_H(f, g+1) = \sum_{j=1}^{f} \kappa_H(j, g). \tag{6'}$$

Using the well-known equation

$$A_g^f = \sum_{j=1}^{f} A_{g-1}^j$$

and the equation (8), we receive the following formula:

$$\kappa_H(f, g) = A_g^{g+f-1}. \tag{12}$$

Substitution of the formula (12) into the equation (11) immediately gives the formula (10). Theorem 1 is proved.

As the transformation ϕ is by definition one-to-one, so the transformation ϕ^{-1} exists, as well.

The rule ϕ is a generalization of the well-known operator *vech*, transforming the upper triangle of a symmetric matrix into the vector.

From Theorem 1 immediately Corollary 1 follows.

COROLLARY 1. The cardinality of the sequence $\mathcal{I}(p,k)$ is A_k^p.

To prove the corollary it is sufficient to use the formula (10), using the values $i_j = p$, $j = 1, \ldots, k$.

Example 1. Let us regard the set of fourth moments of a five-variate random vector $(X_1, X_2, X_3, X_4, X_5)$. The usual way to present the set of moments is to use the moment's matrix (of order $5 \times 5 \times 5 \times 5$). In this matrix, which is usually written with the help of 5×5 blocks, there are 625 entries. But most of the elements of the matrix are superfluous, as almost all moments are represented in this matrix repeatedly. The situation is similar to that of correlation matrix, where all correlation coefficients are represented twice: in the upper and in the lower triangle.

The main practical problem is to eliminate the repeated copies of the moments from the matrix. Using the ordering rule ϕ, where as the dimensionality k the order of moments is taken and as parameter p – the dimensionality p of the initial vector, it is possible to get all different fourth moments of the vector, identified by the indices, as a finite sequence $\mathcal{I}(p,k)$. In Table 1 all different fourth moments of a five-variate random vector are represented as 4-dimensional index-vectors, identifying the indices of the random variables.

Table 1.

$h(I)$	I	$h(I)$	I	$h(I)$	I	$h(I)$	I	$h(I)$	I
1	1111	15	3333	29	4431	43	5331	57	5521
2	2111	16	4111	30	4432	44	5332	58	5522
3	2211	17	4211	31	4433	45	5333	59	5531
4	2221	18	4221	32	4441	46	5411	60	5532
5	2222	19	4222	33	4442	47	5421	61	5533
6	3111	20	4311	34	4443	48	5422	62	5541
7	3211	21	4321	35	4444	49	5431	63	5542
8	3221	22	4322	36	5111	50	5432	64	5543
9	3222	23	4331	37	5211	51	5433	65	5544
10	3311	24	4332	38	5221	52	5443	66	5551
11	3321	25	4333	39	5222	53	5444	67	5552
12	3322	26	4411	40	5311	54	5511	68	5553
13	3331	27	4421	41	5321	55	5521	69	5554
14	3332	28	4422	42	5321	56	5522	70	5555

By the corollary it is easy to calculate the number of different moments $A_5^4 = (8 \cdot 7 \cdot 6 \cdot 5)/4! = 70$. This number is concordant with the data given in Table 1.

To identify the moment $E(X_1 \cdot X_2^2 \cdot X_5)$ we take the index-vector (1,2,2,5), find its representative (5,2,2,1) and calculate the value of $h(I) = A_4^4 + A_3^1 + A_2^1 + A_1^0 + 1 = 35 + 1 + 1 + 0 + 1 = 38$. The result is the rank $h(I)$ of the index-vector $I = (5, 2, 2, 1)$, as we can see in the Table 1.

In similar way the inverse transformation $h \Rightarrow I_h$ can be found. For example, let $h(I) = 50$ be given. As $A_4^4 = 35$ and $A_4^5 = 70$, so $i_1 = 5$. The next step is to compare the difference $50 - 35 = 15$ with the values of A_j^3 $(j = 1, \ldots, 5)$. As $A_3^3 = 10$ and $A_3^4 = 20$, hence $i_2 = 4$. The difference 5 is between the values of $A_2^2 = 3$ and $A_2^3 = 3$ and the difference 5-3 =2 equals to $A_1^2 = A_1^1 + 1$: Hence the index-vector I_{50} has the form (5,4,3,2).

Example 1 demonstrates that the rule ϕ and its inverse ϕ^{-1} are both easily programmable.

4. Repeated ordering

For solving several problems it is necessary to carry out the repeated ordering of index-vectors. For instance, after the moments of order k are estimated, the moments of the estimated empirical moments are needed to estimate. Hence the moments of moments should be calculated. In this case the procedure described in Section 3 can be used, where *as the components i_j of index-vectors* some *index-vectors* should be regarded. In this case the following Theorem 2, generalizing Theorem 1, will hold.

THEOREM 2. Assume $\mathcal{J}$ is a finite set of completely ordered objects i_j, $j = 1, ..., p$, $p = \kappa(\mathcal{J})$. Let $\Psi(p, k)$ be the set of k-variate vectors $I = (i_1, \ldots, i_k)$, where $i_j \in \mathcal{J}$ and the conditions (2) are satisfied, and let $\mathcal{I}(p, k)$ be the lexicographically ordered set $\Psi(p, k)$.

Then the rank $h(I)$ of a vector $(i_1, \ldots, i_k)$ in the sequence $\mathcal{I}(p, k)$ can be calculated by the formula (10).

PROOF is similar to that of Theorem 1.

Example 2.

Let us have the set of empirical fourth moments of a bivariate random vector (X_1, X_2). Using the ordering rule, introduced in Sections 2 – 4, we get the sequence $\mathcal{I}(2, 4)$, consisting of the following five index-vectors (1,1,1,1), (2,1,1,1), (2,2,1,1), (2,2,2,1),

(2,2,2,2). The next problem is to find all 3rd moments of these empirical moments. These moments form a sequence $\mathcal{I}(5,3)$, having the cardinality A_5^7=35. In the following Table 2 the double index-vectors, consisting of index-vectors as components, are given.

Notice, that by the usual algorithms of calculation of moments, we should have received the matrix of moments, consisting of $(2^4)^3$= 4096 entries instead of 35, given in Table 2.

Table 2.

$h(I)$	$I=(I_1,I_2,I_3)$	18	(2221),(2221),(2111)
1	(1111),(1111),(1111)	19	(2221),(2221),(2211)
2	(2111),(1111),(1111)	20	(2221),(2221),(2221)
3	(2111),(2111),(1111)	21	(2222),(1111),(1111)
4	(2111),(2111),(2111)	22	(2222),(2111),(1111)
5	(2211),(1111),(1111)	23	(2222),(2111),(2111)
6	(2211),(2111),(1111)	24	(2222),(2211),(1111)
7	(2211),(2111),(2111)	25	(2222),(2211),(2111)
8	(2211),(2211),(1111)	26	(2222),(2211),(2211)
9	(2211),(2211),(2111)	27	(2222),(2221),(1111)
10	(2221),(1111),(1111)	28	(2222),(2221),(2111)
11	(2221),(2111),(1111)	29	(2222),(2221),(2211)
12	(2221),(2111),(1111)	30	(2222),(2221),(2221)
13	(2221),(2111),(2111)	31	(2222),(2222),(1111)
14	(2221),(2211),(1111)	32	(2222),(2222),(2111)
15	(2221),(2211),(2111)	33	(2222),(2222),(2211)
16	(2221),(2211),(2211)	34	(2222),(2222),(2221)
17	(2221),(2221),(1111)	35	(2222),(2222),(2222)

It is also easy to see, that the result is different from the case, if all components were used in the same index-vector of length $k = 3 \cdot 4$. In the last case the cardinality of the sequence was A_2^{12}= 13, that means, part of the moments, represented in Table 2, were missing.

5. Classification of higher moments of a random vector

In the set $\Psi(p,k)$ there are different types of fourth moments, depending on the number of variables, included into the moment and the degrees of them. To describe the higher moments of a random vector it is convenient to use the concept of the *partition of a number*, see e.g. Andrews (1980).

DEFINITION 5. Let n be a natural number. A set of natural numbers $Q = \{q_1, \ldots, q_v\}$, satisfying the condition

$$n = \sum_{j=1}^{v} q_j$$

is a *partition of the natural number* n, the addendae q_j are said to be the *parts* of n, $v = \kappa(Q)$ is the *cardinality of the partition* Q.

Usually the set Q is represented as a non-decreasingly ordered set of parts. As the parts have the values $1, 2, \ldots, n$, so there exists an equivalent representation of the partition via its characteristic numbers a_i, $i = 1, \ldots, n$ in the following sense.

DEFINITION 6. Let $Q = \{q_1, \ldots, q_v\}$ be a partition of a natural number n. Let a_j denote the number of repetitions of the integer j in the set Q, i.d., the following equation must be valid:

$$n = \sum_{j=1}^{n} j \cdot a_j.$$

The numbers a_j are said to be the *characteristic numbers* of the partition Q, and the vector $a = (a_1, \ldots a_n)$ is its *characteristic vector.*

Let us regard the set of moments of order k of a p-dimensional random vector, $k \leq p$, and introduce the following classification, using the concept of the *partition of a number.*

DEFINITION 7. Let $(X_1, \ldots, X_p)$ be a given random vector and $Q = (q_1, \ldots, q_v)$ a *partition of a natural number* $k, k \leq p$. Then we say that the moment

$$E(X_{i_1}^{q_1} \cdot \ldots \cdot X_{i_v}^{q_v}) \tag{13}$$

belongs to the *type* Q, if $i_1, \ldots, i_v$ are different integers from the set $\{1, \ldots, p\}$.

The following problem is to clear up the number $\kappa_p(Q)$ of moments of a type Q in the sequence $\mathcal{I}(p, k)$. The answer to this problem will be given in the following theorem.

THEOREM 3. Let $\mathcal{I}(p, k)$ be a sequence of all moments of the order k of a p-variate random vector, and let Q be a partition of the number k, having the characteristic vector $a = (a_1, \ldots, a_k)$.

Then the number of moments of the type Q equals to the following expression:

$$\kappa_k(Q) = \prod_{j=1}^{k} C_{d_j}^{a_j}, \tag{14}$$

where:

$$d_j = p - b_{j-1}, \;\; b_j = \sum_{f=1}^{j} a_j, \; b_0 = 0; \; j = 1, \ldots, k$$

and $C_n^m = n!/(m! \cdot (n-m)!)$.

PROOF is evident. Let r be the first (smallest) integer having a non-zero characteristic number a_r. Then there are C_p^r different ways to fix the indices of the variables X_i of the a_r-th degree in the expression (13) of the k-th moment. For the following term having non-zero degree there remains $p - a_r$ possibilities, etc.

Example 3. Let us regard the set of the fourth moments of a five-variate index-vector, given in Example 1. As the number 4 (the order of moments k) has 5 different partitions ([4], [3,1], [2,2], [2,1,1], [1,1,1,1]), so there are five different types of the fourth moments, see Table 1. To the partition [4] all marginal moments correspond, to the partition [2,2] the moments $E(X^2 \cdot Y^2)$, etc.

Let us calculate the cardinalities of different fourth moments of all types. At first, let us write the characteristic vectors of all partitions:

$[4] \Rightarrow (0,0,0,1)$; $[3,1] \Rightarrow (1,0,1,0)$; $[2,2] \Rightarrow (0,2,0,0)$;
$[2,1,1] \Rightarrow (2,1,0,0)$; $[1,1,1,1] \Rightarrow (4,0,0,0)$;

The cardinalities are calculated using the formula (14):

$\kappa\{[4]\} = C_5^1 = 5$; $\kappa\{[3][1]\} = C_5^1 \cdot C_4^1 = 20$;
$\kappa\{[2][2]\} = C_5^2 = 10$; $\kappa\{[2][1][1]\} = C_5^1 \cdot C_4^2 = 30$;
$\kappa\{[1][1][1][1]\} = C_5^4 = 5$;

It is easy to check that the results of the calculation fit with the data in Table 1.

For some types of moments it is quite easy to find the rule for determining their ranks in the sequence $\mathcal{I}(p,k)$. E.g., the marginal moment, corresponding to index j, has the rank

$$h(I) = A_k^j, \; j = 1, \ldots, p.$$

6. Binary partition of index-vectors

Often in the practical multivariate statistical analysis the following problem arises. Let a random vector $\vec{X} = (X_1, \cdots, X_p)$ be given. From $\vec{X}$ a subvector $(X_{i_1}, \cdots X_{i_q})$ can be extracted. To formalize this procedure it is convenient to use the index-vector:

$$I = (i_1, \cdots, i_q),$$

satisfying the following conditions:

1^0 all components i_j of I are natural numbers, $i_j \leq p$;

2^0 $i_j < i_{j+1}$, $j = 1, \cdots, q-1$.

3^0 $q \leq p$.

Here q is the *cardinality of the index-vector* I, $q = \kappa(I)$.

Another class of problems is connected with the partition of a given random vector $\vec{X}$ into two subvectors. For solving this problem a pair of complementary index-vectors I and I^c can be used, where I^c can be defined as the index-vector, satisfying the conditions $1^0 - 2^0$ and consisting of indices:

$$\{j_1, \cdots, j_r\} = \{1, \cdots, p\} - \{i_1, \cdots, i_q\}.$$

It is evident, that the number of components r of the index-vector I^c equals to $r = p - q$.

In the following we will use the symbol I (or I^c) of an index-vector for denoting the set of its components, as well:

$$I = \{i_1, \cdots, i_q\}.$$

We will use the term *order* for the *cardinality* p of the initial index-set $(i_1, \ldots, i_k)$.

DEFINITION 8. Let the integer p be fixed. Every pair of complementary index-vectors (I, I^c), satisfying the conditions $1^0 - 3^0$, defines a *binary partition or shortly B-partition* B_p of the index-vector $(1, \ldots, p)$. Here p is said to be the *order* of the partition.

As the index-vector I^c is uniquely defined by the index-vector I, we will in future identify the B-partition by the (first) index-vector I and use the denotation $B_p(I)$.

A. Our following task is to find the rule for construction of binary partitions B_p. This task can be solved easily with using the *indicators*, defined in the following way:

DEFINITION 9. A p-variate vector $L = (l_1, \ldots, l_p)$, is said to be an *indicator of the partition* $B_p(I)$, if

$$l_i = \begin{cases} 1, \text{if } i \in I, \\ 2, \text{if } i \in I^c. \end{cases} \tag{15}$$

Using the concept of indicator, it is easy to construct *all* B_p*-partitions* with the help of *binary numbers as indicators.* From here it follows that the cardinality of the set of p-variate indicators should be 2^p. As for binary numbers, consisting of 1's only and 2's only, no partition corresponds, hence the cardinality of the set $\mathcal{B}_p$ is $2^p - 2$.

As in the partition $B_p(I)$ the parts I and I^c are, in principal, symmetric, so sometimes it is important to guarantee the uniqueness of an partition with the help of an additional condition. The simpliest way is to demand that

$$i_1 \in I. \tag{16}$$

Let us denote the B_p-partitions, satisfying the condition (16) by $B'_p(I)$ and say that these are B'_p-partitions. Let the set of B'_p-partitions of order p be $\mathcal{B}'_p$. It is evident, that there exists one-to-one correspondence between all B_{p-1}-partitions of order $p-1$ and B'_p-partitions of order p, realized on the level of indicators $L = (l_1, \ldots, l_{p-1})$ and $L' = (l'_1, \ldots, l'_p)$,

$$l'^f_j = \begin{cases} 1, \text{ if } j = 1, \\ l^f_{j-1}, \text{ if } j = 2, \ldots, p, \end{cases}$$

$f = 1, \ldots, \kappa(\mathcal{B}'_p)$. Hence $\kappa(\mathcal{B}'_p) = \kappa(\mathcal{B}_{p-1}) = 2^{p-1}$, where also the partition $(I^0_p, \emptyset)$ is included into the set $\mathcal{B}'_p$.

C. A more complicated task is to find specially defined subsets of all B-partitions. E.g., in regression analysis it is convenient to find all subsets of explanatory variables, having given cardinality q. For solving problems of this kind we will introduce the concept of the *cardinality* of a B_p-partition in the following way:

DEFINITION 10. Let us have a B_p-partition $B_p(I)$. We say that its *cardinality* $\kappa(B_p)$ is q, if $\kappa(I) = q$.

We will denote the B-partition (B'-partition) of the order p and of cardinality q by B^q_p (B'^q_p, correspondingly).

Our following step is to give an algorithm for the construction of B^q_p-partitions of fixed order and cardinality.

Algorithm 1. Construction of indicators of B_p^q-partitions. Let the integers p (order of the index-vector) and q (cardinality) be given, $1 \leq q \leq p$.

Step 1. The indicator $L^1 = (l_1^1, \ldots, l_p^1)$ is defined in the following way:

$$l_j^1 = \begin{cases} 1, \text{if } j = 1, \ldots, q, \\ 2, \text{if } j = q+1, \ldots, p. \end{cases}$$

Step 2. Let us have constructed f indicators ($f \geq 1$), and let the indicator L_f have the following form:

$$(l_1^f, \ldots, l_{p-a-b}^f, 1, \overbrace{2, \ldots, 2}^{a}, \overbrace{1, \ldots, 1}^{b}),$$

where $1 \leq a \leq p-q$, $0 \leq b \leq q$. Let us notice, that the first $p-a-b-1$ components of the index-vector L_f have no influence on the construction. The construction of the next index-vector L_{f+1} depends on the value of b:

If $b = q$, then Step 4.

If $b = 0$, then $_{f+1}$ will be defined in the following way:

$$l_j^{f+1} = \begin{cases} 2, \text{ if } j = p-a, \\ 1, \text{ if } j = p-a+1, \\ l_j^f \text{ otherwise.} \end{cases}$$

If $0 < b < q$, then

$$l_j^{f+1} = \begin{cases} l_j^f, \text{ if} j = 1, \ldots, p-a-b-1, \\ 2, \text{ if } j = p-a-b, \\ 1, \text{ if } j = p-a-b+1, \cdots, p-a+1, \\ 2, \text{ if } j = p-a+2, \cdots, p. \end{cases}$$

As it follows, the index-vector L_{f+1} has in this case the following form:

$$(l_1^{f+1}, \ldots, i_{p-a-b}^{f+1}, 2, \overbrace{1, \ldots, 1}^{b+1}, \overbrace{2, \ldots, 2}^{a-1})$$

L_{f+1} is defined.

Step 3. Take

$f := f + 1$ and repeat Step 2.

Step 4.. End. The sequence $\{L_f\}$ of all indicators of B_p^q-partitions of order p and cardinality q is completed.

Example 4. Let us demonstrate the sequence L_f in the case $p = 7$, $q = 4$, see Table 3.

Table 3.

f	I_f	12	(1211212)	24	(2112112)
1	(1111222)	13	(1211221)	25	(2112121)
2	(1112122)	14	(1212112)	26	(2112211)
3	(1112212)	15	(1212121)	27	(2121112)
4	(1112221)	16	(1212211)	28	(2121121)
5	(1121122)	17	(1221112)	29	(2121211)
6	(1121212)	18	(1221121)	30	(2122111)
7	(1121221)	19	(1221211)	31	(2211112)
8	(1122112)	20	(1222111)	32	(2211121)
9	(1122121)	21	(2111122)	33	(2211211)
10	(1122211)	22	(2111212)	34	(2212111)
11	(1211122)	23	(2111221)	35	(2221111)

From Algotithm 1 the following corollaries can be deduced:

COROLLARY 2. In the sequence L_f, constructed by Algorithm 1, all indicators are lexicographically ordered.

COROLLARY 3. The cardinality $\kappa\{L_f$ of the sequence, constructed by Algorithm 1, is C_p^q.

C. Construction of $B_p'^q$-partitions. From Algorithm 1 for building the indicators of B_p^q-partitions it is easy to get an algorithm for building the indicators of $B_p'^q$-partitions, which satisfy the condition (16) and have the parameters p (order) and q (cardinality).

Algorithm 2. Construction of indicators of $B_p'^q$-partitions. Let the integers p (the order) and q (the cardinality) be given, $1 \le q \le p$.

If $q = 1$, then by the condition (16) exactly one partition, corresponding to the indicator $(1, 2, \ldots, 2)$, exists.

Let us regard the case $q > 1$ and define $q' := q-1, p' := p-1$.

Step 1. Calculate $F = C_{p'}^{q'}$. Use Algorithm 1 with parameters p' and q'.

Step 2. Let L_f be the f-th indicator generated by the Algorithm 1. Then

$$L'_f = (1, l_1, ^f, \ldots, l_{p'})$$

is the indicator of $B_p'^q$-partition.

Step 3. If $f = F$, then Step 4.

If $f < F$, then $f : f + 1$, repeat Step 2.

Step 4. End.

COROLLARY 4. The cardinality of the sequence $\{L'_f\}$ is $C_{p'}^{q'}$.

Example 5. Let us construct the sequence $\{L'_f\}$ of indicators for $B_p'^q$-partitions having the order $p = 7$ and the cardinality $q = 4$.

It is easy to see, that the 20 first indicators from Table 3 are exactly the indicators L'_f needed.

D. Construction of indicators of all B'_p-partitions.

Using Algorithm 2 repeatedly, taking $q = 1, \cdots, p$ all B'_p-partitions can be received. The cardinality of the set of all B'_p-partitions is

$$\kappa(\mathcal{B}'_p) = \sum_{q'=0}^{p'} C_{p'}^{q'} = 2^{p'} = 2^{p-1}. \tag{17}$$

Example 6. Let us construct the indicators of all B'_5-partitions, see Table 4.

Table 4.

f	I'_f	f	I'_f	f	I'_f	f	I'_f
1	(12222)	5	(12221)	9	(12112)	13	(11121)
2	(11222)	6	(11122)	10	(12121)	14	(11211)
3	(12122)	7	(11212)	11	(12211)	15	(12111)
4	(12212)	8	(11221)	12	(11112)	16	(11111)

We see, that there is one indicator corresponding to cardinality, equal to 1, a subsequence, consisting of 4 indicators, corresponding to cardinality, equal to 2 etc. The last indicator corresponds to the trivial B'_p-partition, where $I = \{1, \ldots, p\}$, $I^c = \emptyset$. Every subsequence is lexicographically ordered, but not the compound sequence $\{L'_f\}$.

7. Arbitrary partition of an index-vector

Let us regard the following problem. The index-set $I_p^0 = \{1, \ldots, p\}$ and a natural number s, $1 < s \leq p$ are given. The problem is to find all possible partitions P_p^s of the set I_p^0 into not more than s non-overlapping parts.

This task must be solved for getting all quasi-maximal distributions, having not more than s independent (higher) marginal

distributions, or for constructing all quasi-positive correlation matrices, consisting of not more than q blocks (of non-correlated variables), see Tiit (1986).

Every partition can be characterized by an index-vector $I^j = (i_1^j, \ldots, i_{v_j}^j)$, $j = 1, \cdots, s$, $\sum_{j=1}^{s} v_j = p$.

To solve this problem it is useful to introduce the concept of the *s-indicator*, generalizing the indicator (see Definition 9) in the following way:

DEFINITION 11. A p-variate vector $L = (l_1, \ldots, l_p)$, satisfying the condition $1 \leq i_j \leq s$, $j = 1, \ldots, p$, is said to be an *s-indicator* of a s-partition, if

$$l_i = j \leftrightarrow i \in I_j, \; i = 1, \ldots, p; \; j = 1, \ldots, s. \tag{18}$$

To guarantee the uniqueness of the s-partition constructed by an s-indicator, it is useful to introduce the conditions, generalizing the condition (16) in the following way:

$$i_1^j < i_1^{j+1}, \; j = 1, \ldots, s-1. \tag{19}$$

Let us give an algorithm for building all q-indicators (18) of P_p^q-partitions, $q = 1, \ldots, s$, satisfying the conditions (19).

Algorithm 3. Construction of indicators of P_p^q-partitions.

Step 1. Take $f = 1$,

$L^1 = (1, \ldots, 1)$.

Step 2. Let f indicators be constructed. Define the values $m_i, i = 1, \ldots, p$ in the following way:

$$m_1 = 1, \; m_i = \min(s, \max_{1 \leq j \leq i-1} l_j^f + 1), \; i = 2, \ldots, p.$$

If $l_p^f < m_p$, then the indicator L^{f+1} will be constructed by the following formula:

$$l_j^{f+1} = \begin{cases} l_j^f, \text{ if } j = 1, \ldots, p-1, \\ l_p^f + 1, \text{ if } j = p. \end{cases}$$

If $l_p^f = m_p$, then find the first index q satisfying the condition:

$$l_{p-q}^f < m_{p-q}, \; q < p. \tag{20}$$

If there does not exist any integer q, satisfying the copndition (20), then Step 4.

If there exists such index q, then define the indicator L^{f+1}:

$$l_j^{f+1} = \begin{cases} l_j^f, \text{ if } j = 1, \ldots, p-q-1, \\ i_{p-q}^f + 1, \text{ if } j = p-q, \\ 1, \text{ if} j = p-q+1, \ldots, p. \end{cases}$$

Step 3. Take $f := f + 1$ and repeat Step 2.
Step 4. End, the sequence $\{L_f\}$ is completed.

COROLLARY 5. From the construction it follows that the sequence $\{L^f\}$ is lexicographically ordered.

Let us use the notation $\mathcal{P}_p^s$ for the set of all P_p^s-partitions and $\mathcal{P}_p$ for the set of all p-partitions.

COROLLARY 6. Taking $s = p$ in Algorithm 3 we get all indicators of the partitions of the set $\mathcal{P}_p$ in lexicographical order.

Example 7. Let us construct the sequence of all indicators of the P_5^3-partitions (see Table 5) and all indicators of the P_5-partitions (see Table 6).

Table 5.

f	I_f	11	(11223)	22	(12132)	33	(12311)
1	(11111)	12	(11231)	23	(12133)	34	(12312)
2	(11112)	13	(11232)	24	(12211)	35	(12313)
3	(11121)	14	(11233)	25	(12212)	36	(12321)
4	(11122)	15	(12111)	26	(12213)	37	(12322)
5	(11123)	16	(12112)	27	(12221)	38	(12323)
6	(11211)	17	(12113)	28	(12222)	39	(12331)
7	(11212)	18	(12121)	29	(12223)	40	(12332)
8	(11213)	19	(12122)	30	(12231)	41	(12333)
9	(11221)	20	(12123)	31	(12232)		
10	(11222)	21	(12131)	32	(12233)		

Using the value $q = 5$ we get all P_5-partitions. In Table 6 there are all indicators, contained in Table 5 plus some indicators, containing the indices 4 and 5, additionally.

Table 6.

1	(11111)	14	(11233)	27	(12212)	40	(12321)
2	(11112)	15	(11234)	28	(12213)	41	(12322)
3	(11121)	16	(12111)	29	(12221)	42	(12323)
4	(11122)	17	(12112)	30	(12222)	43	(12324)
5	(11123)	18	(12113)	31	(12223)	44	(12331)
6	(11211)	19	(12121)	32	(12231)	45	(12332)
7	(11212)	20	(12122)	33	(12232)	46	(12333)
8	(11213)	21	(12123)	34	(12233)	47	(12334)
9	(11221)	22	(12131)	35	(12234)	48	(12341)
10	(11222)	23	(12132)	36	(12311)	49	(12342)
11	(11223)	24	(12133)	37	(12312)	50	(12343)
12	(11231)	25	(12134)	38	(12313)	51	(12344)
13	(11232)	26	(12211)	39	(12321)	52	(12345)

8. Multiple partitions

Using the partition rules (Algorithms 1 - 3) for all index-vectors I^j , characterizing the parts of a partition of a given index-vector I_p^0, it is easy to get several types of *multiple partitions*.

One task, using the double partition, is the construction of quasi-extremal distributions (correlation matrices), see Tiit (1986). First, the partitions of the initial index-vector into independent subvectors must be found. Second, all extremal distributions of defined subvectors must be constructed, e.g., all B-partitions of each part should be found. This procedure is illustrated in the following example.

Example 7. Let us construct all B-partitions of all 4-partitions. In the Table 7 all indicators of these double partitions are given, where the parts of the 4-partitions are denoted with integers from 1 to 4, and the parts of the B-partition - with signs before these integers, where the sign + (omitted) corresponds to the first index-vector I and the sign − to the second index-vector I^c.

Table 7.

f	L^f	17	1,1,2,2	34	1,1,1,−1
1	1,1,1,1	18	1,−1,2,2	35	1,2,−2,1
2	1,1,1,−1	19	1,1,2,−2	36	1,2,−2,−1
3	1,1,−1,1	20	1,−1,2,−2	37	1,2,2,2
4	1,1,−1,−1	21	1,1,2,3	38	1,2,−2,2
5	1,−1,1,1	22	1,−1,2,3	39	1,2,−2,2
6	1,−1,1,−1	23	1,2,1,1	40	1,2,−2,−2
7	1,−1,−1,1	24	1,2,1,−1	41	1,2,2,3
8	1,−1,−1,−1	25	1,2,−1,1	42	1,2,−2,3
9	1,1,1,2	26	1,2,−1,−1	43	1,2,3,1
10	1,1,−1,2	27	1,2,1,2	44	1,2,3,−1
11	1,1,−1,2	28	1,2,−1,2	45	1,2,3,2
12	1,−1,−1,2	29	1,2,1,−2	46	1,2,3,−1
13	1,1,2,1	30	1,2,−1.−2	47	1,2,3,3
14	1,1,2,−1	31	1,2,1,3	48	1,2,3,−3
15	1,−1,2,1	32	1,2,−1,3	49	1,2,3,4
16	1,−1,2,−1	33	1,2,2,1		

To explain the using of the indicators from Table 7 for construction of quasi-extremal distributions and correlation matrices let us take one indicator, e.g. L_{30}. To this indicator (1,2,-1,-2) there corresponds a distribution, where the first and the third components are independent from the second and the fourth ones. At the same time, the first and the third components are negatively intercorrelated, and the same is true about the second and the fourth components, as well.

In the case of equal and symmetrical univariate marginal distributions the correlation matrix, corresponding to the double partition, described by the indicator L_{30}, is the following:

$$\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

References.

Andrews, G.(1980) Partition Theory.

Fisher, R.A. (1929) Proc. Lond. Math. Soc. (2), **30**, 199 - 238.

Kotz,S., Johnson, N.L. (1981) Enzyclopedia of Statistical Sciences, J.Wiley & Sons, N.York.

Tiit, E.-M. (1984) Construction of discrete multivariate distributions having given moments. Discrete analogue of the normal distribution. In: Proceedings of Computing Centre of University of Tartu,**51**, 142 - 165 (in Russian).

Tiit, E.-M. (1986) Random vectors with given arbitrary marginals and given correlation matrix. - Acta et Commentationes Universitatis Tartuensis, **733**, 14 - 38.

Tiit, E.-M. (1988) Unbiased and k-biased estimations of entire rational functions of moments. - Acta et Commentationes Universitatis Tartuensis, **798**, 3 - 18.

Tiit, E.-M. (1992) Extremal multivariate distributions having given discrete marginals. Acta et Commentationes Universitatis Tartuensis, **942**, 94 - 113.

Received January, 1994

Indeks-vektorite ja liigenduste kasutamine mitmemõõtmelises statistilises analüüsis

Ene-Margit Tiit, Sirli Tammet

Kokkuvõte

Käesolevas artiklis vaadeldakse mitmemõõtmelises statistilises analüüsis sageli kasutatavat *indeks-vektori* mõistet, mis hästi sobib juhusliku vektori alamvektori identifitseerimiseks. Rakendades indeks-vektorile *liigendusi*, saame identifitseerida juhuslike suuruste hulga jaotamise osahulkadeks. Artiklis tõestatakse mõned teoreemid indeks-vektorite leksikograafiliselt järjestatud jada kohta, millest on kasu juhusliku vektori momentide hulga kirjeldamisel (nn Fisheri momentide momentide probleem), samuti tuletatakse algoritme liigenduste moodustamiseks.

Acta et Commentationes Universitatis Tartuensis, **968** , 1994, pp. 77-97.

The Convex-extremal Decomposition of Correlation Matrix. An Application in Anthropometrical Research

Ene-Margit Tiit, Maie Thetloff

Summary. In the paper the extremal and quasi-extremal distributions and corresponding correlation matrices (consisting of ones and minus ones, or ones, zeros and minus ones) are introduced. Using the extremal and quasiextremal correlation matrices the convex-extremal decomposition of a given correlation matrix is defined, and compared with factor analysis. Several examples and a practical application in physical anthropology are given.

Key words: *correlation matrix, factor analysis, convex decomposition, mixture, extremal distribution*

1. Set-up of the problem

For a long time the factor analysis has been the most popular methods in multivariate statistics to establish and describe the *dependence structure* of a high-dimensonal data-set, using the decomposition of a given correlation matrix by its eigenvalues and eigenvectors, so-called *factor decomposition*. Here we propose another methodology – the convex decomposition of a correlation matrix by the extremal correlation matrices, or shortly *the convex-extremal decomposition* of a correlation matrix. We will also analyse the common and different features of the two decompositions. Some comparative examples of both analyses will be given.

As the extremal correlation matrices are in many cases *standard* in the sense that they do not depend on concrete data-set, so it is reasonable to use the convex-extremal decomposition for describing the *differences* between data-sets or *changes in time*. In

the last part of the article an example of the usage of the convex-extremal decomposition in the description of the change of the anthropometrical structure of schoolchildren will be given.

2. The underlying models

Both models use the correlation matrix as the initial information. Hence, they describe the correlative dependencies and for both methods the best theoretical model is the normally distributed population. Nevertheless, for the convex-extremal analysis arbitrary multivariate distribution having equal symmetrical marginals can be used.

In practical applications the most convenient result is the *approximate model*, which describes only a part of the information (in the sense of dependencies) of the initial distribution.

In both cases the building of the model means the estimation of parameters of the model, more precisely – the *calculation of the coefficients of the decomposition* of a correlation matrix by some other matrices.

As it is well-known, the factor analysis uses the decomposition of the given correlation matrix R of order k by the eigenvectors and eigenvalues:

$$R = H\Lambda H', \tag{1}$$

where H is the matrix of eigenvectors $h_g = (h_1^g, \ldots, h_k^g)$, $g = 1, \ldots, k$, Λ the matrix of non-increasingly ordered eigenvalues, $\Lambda = diag(\lambda_1, \ldots, \lambda_k)$ and k the order of matrix R. In factor analysis usually instead of matrix H the block of H, consisting of q first columns, $q < k$, is used.

The extremal decomposition of the correlation matrix has the following form:

$$R = \sum_{g=1}^{h} w_g R_{I_g}, \tag{2}$$

where the weights w_g must fulfill the following conditions:

$$w_g \geq 0, \quad \sum_{g=1}^{h} w_g = 1. \tag{3}$$

and the number h of components of the decomposition fulfills the condition

$$1 \leq h \leq k(k-1)/2 + 1.$$

The matrices R_{I_q} belong either to the set of extremal correlation matrices (see Tiit, 1984, 1986, 1992, Kotz, Tiit, 1992) or to the set of quasi-extremal matrices (see Tiit, 1986). In the case of equal and symmetric, e.g. normal population the extremal correlation matrices consist of ones and minus ones only. The number of extremal correlation matrices is 2^{k-1}. The quasi-extremal correlation matrices consist of zeros, ones and minus ones, and their number is much larger, but, of course, always finite.

The factor decomposition of a correlation matrix can be interpreted as *the linear transformation of variables* – the initial variables are expressed as linear combinations of (small number) factors, having some good standard properties (standardization, orthogonality).

The convex-extremal decomposition of a correlation matrix can be interpreted as a *mixture of exteremal matrices* and hence *the initial distribution can be expressed as a mixture of extremal distributions* (see Tiit, 1984, 1986, 1992), having the same univariate marginal distributions as the initial distribution. In the next paragraph the properties of the extremal correlation matrices will be regarded more closely.

3. The extremal correlation matrix and the convex decomposition

Let $X_1, \ldots, X_k$ be given random variables having the same symmetrical distribution P_o. Then the maximal and minimal possible correlation coefficients between the variables X_i and X_j are equal to 1 and -1, correspondingly, $i, j = 1, \ldots, k$.

In the case $k = 2$ the concept of *minimal* and *maximal* distributions was introduced by Hoeffding (see Hoeffding, 1940) and Frechet (see Frechet, 1951).

An extremal correlation matrix is defined with the help of a binary partition of the set of initial variables, as it follows from the forthcoming definitions, (see Tiit, Tammet, 1994).

DEFINITION 1. Let I^o be the initial index-set $\{1, 2, \ldots, k\}$. Say (I, I^c) is *a partition (more precisely binary partition) of the index-set* I^o, if the subsets $I = \{i_1, \ldots, i_q\}$ and $I^c = \{j_1, \ldots, j_s\}$ of I^o are non-overlapping and fulfill the following conditions:

$$i_1 = 1, \quad I \cup I^c = I^o. \tag{4}$$

Hence the cardinalities q and s of the subsets fulfill the fol-

lowing conditions:

$$1 \leq q \leq k,\ 0 \leq s = k - q \leq k - 1.$$

From the definition it follows that as a special case the trivial partition $(I^o, \emptyset)$ can be regarded. Then the common number of different partitions of the index-set I^o is 2^{k-1}.

DEFINITION 2. Let (I, I^c) be a partition. The correlation matrix $R_I = (r_{ij}^I)$, defined in the following way

$$r_{ij}^I = \begin{cases} 1, \text{ if } i \in I \text{ and } j \in I \text{ or } i \in I^c \text{ and } j \in I^c \\ -1, \text{ if } i \in I \text{ and } j \in I^c \text{ or } i \in I^c \text{ and } j \in I \end{cases} \tag{5}$$

is said to be *the extremal correlation matrix, corresponding to the partition* (I, I^c).

We shall use the symbol $\mathbf{R}(k)$ for denotation the set of all extremal correlation matrices of order k.

From here it follows that between all variables, eg. X_i and X_j should be an exact linear dependency:

$$X_i = a_{ij} + b_{ij} X_j,\ i, j = 1, \ldots, k. \tag{6}$$

where the sign of regression coefficient b_{ij} is determined by the direction of correlation, e.g. by the sign of r_{ij}.

The correlation matrix corresponding to the partition $(I^o, \emptyset)$, is so-called *maximal correlation matrix*. All its elements equal to the maximal correlation coefficients, which equal in the case of identical marginals to one. Hence, in the formula (6) all regression coefficients b_{ij} are strictly positive.

DEFINITION 3. Let R be a given correlation matrix of order k. If there exist such coefficients w_g, $g = 1, \ldots, h$, fulfilling the conditions (3), that the equation (2) holds, where $R_{I_g} \in \mathbf{R}(k)$, we will say that *for the correlation matrix R there exists a convex decomposition by the extremal correlation matrices.*

If the decomposition (2) exists, it will be easy to show that the decomposition, having the number of terms h, $h < k(k-1)/2+1$, exists as well.

In future we will need the well-known concept of *orthogonality* in the case of quadratic matrices, as it will be given in the following:

DEFINITION 4. Let A and B be quadratic matrices of order k. They are said to be *orthogonal*, if the following condition holds:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}b_{ij} = 0.$$

4. Comparison of factor decomposition and convex-extremal decomposition

Let R be a correlation matrix. Let us rewrite the factor decomposition (1):

$$R = \sum_{g=1}^{k} \lambda_g h_g h_g' = \sum_{g=1}^{k} v_g Q_g, \tag{7}$$

where $v_g = \lambda_g/k$ and the matrix $Q_g = kh_g h_g'$ is defined by its elements:

$$q_{ij}^g = kh_i^g h_j^g. \tag{8}$$

About the decomposition the following lemmas hold:

LEMMA 1. The coefficients v_g fulfill the conditions (3) .

Proof follows immediately from the well-known properties of eigenvalues of the correlation matrix:

$$\lambda_j \geq 0, \ \sum_{j=1}^{k} \lambda_j/k = 1.$$

LEMMA 2. The matrix Q_g, $g = 1, \ldots, k$, defined by the formula (7), has the following properties:

1) Q_g is positively defined;

2) $\mathrm{Tr}Q_g = k$;

3) If $g \neq f$, then Q_g and Q_f are orthogonal.

Proof follows from the definition of matrices Q_g and from the properties of eigenvectors of a correlation matrix, as it will be shown in the following three equations:

$$xQx' = k\sum_{i=1}^{k}\sum_{j=1}^{k} x_i h_i^g h_j^g x_j = \Big(\sum_{i=1}^{k} x_i h_i^g\Big)^2 \geq 0;$$

$$\mathrm{Tr}\, Q_g = \sum_{i=1}^{k} q_{ii}^g = k \sum_{i=1}^{k} (h_i^g)^2 = k;$$

$$\sum_{i=1}^{k}\sum_{j=1}^{k} q_{ij}^g q_{ij}^f \;=\; k^2 \Big(\sum_{i=1}^{k} h_i^g h_i^f\Big)^2 \;=\; 0.$$

LEMMA 3. The matrix Q_g, defined by the formula (7), is a correlation matrix if and only if all components of the eigenvector h_g have the equal absolute values.

Proof. I. Let us suppose that $h_j^g = \max_{1\le i\le k}(h_i^g)$ and there exists an index l so that the inequality $\mid h_l^g \mid < \mid h_j^g \mid$ is valid. Let us use the following denotation: $\bar{h}(2) = 1/k \sum_{i=1}^{k} (h_i^g)^2$. Then $\bar{h}(2)$ is the average of the squares of the components of the eigenvector h_g, and from the properties of the average we have the following inequality:

$$\mid h_j^g \mid^2 > \; 1/k \sum_{i=1}^{k} (h_i^g)^2 = 1/k,$$

hence $\mid q_{jj}^g \mid > 1$ and Q_g cannot be a correlation matrix.

II. Let $\mid h_i^g \mid = h$ for $i = 1, \ldots, k$, then $h = 1/k$ and $\mid q_{ij}^g \mid = 1$ for $i = 1, \ldots, k$. From Lemma 2 it follows that in this case Q_g is a correlation matrix.

COROLLARY 4. If all components of an eigenvector h_g of a correlation matrix R have the equal absolute values, then the matrix Q_g, defined by the formula (8), is an extremal correlation matrix. The partition (I, I^c) corresponding to Q_g is defined in the following way:

$$i \in \begin{cases} I_g, & \text{if } h_i^g > 0, \\ I_g^c, & \text{if } h_i^g < 0. \end{cases}$$

From lemmas 1 - 3 and the corollary 4 the following Theorem 5 can be deduced:

THEOREM 5. A. The factor decomposition of a correlation matrix R is in the same time the convex decomposition of the given correlation matrix R by the extremal matrices, if all components of every eigenvector, used in the decomposition, have correspondingly equal absolute values:

$$\mid h_i^g \mid = h_g, \; i = 1, \ldots, k, \; g = 1, \ldots, q.$$

B. The convex decomposition of a correlation matrix R by its extremal correlation matrices R_{I_j} is in the same time the factor decomposition, if all extremal correlation matrices used in the decomposition are orthogonal.

From here it follows, that in this case the number of components q of the decomposition cannot be more than k.

5. Example 1. Comparison of decompositions in the case of an artifical correlation matrix

Let us regard the correlation matrix

$$\begin{pmatrix} 1 & 0.6 & 0.4 & 0 \\ 0.6 & 1 & 0 & 0.4 \\ 0.4 & 0 & 1 & 0.6 \\ 0 & 0.4 & 0.6 & 1 \end{pmatrix}.$$

It is easy to check that this matrix has the following eigenvalues: $\lambda_1 = 2$, $\lambda_2 = 1.2, \lambda_3 = 0.8; \lambda_4 = 0$ and eigenvectors: $h_1 = (0.5, 0.5, 0.5, 0.5)$, $h_2 = (0.5, 0.5, -0.5, -0.5)$, $h_3 = (0.5, -0.5, 0.5, -0.5)$. The fourth eigenvector is not defined, as $\lambda_4 = 0$.

The full factor matrix consists of three columns that are equal to the product $\sqrt{\lambda_g} h_g$ and has the following form:

$$\begin{pmatrix} 0.707 & 0.548 & 0.447 \\ 0.707 & 0.548 & -0.447 \\ 0.707 & -0.548 & 0.447 \\ 0.707 & -0.548 & -0.447 \end{pmatrix}.$$

As all components of the eigenvectors have the equal absolute values, so the eigenvectors define the extremal correlation matrices. Corresponding partitions are the following (see Corollary 1):

$$I_1 = \{1, 2, 3, 4\},\ I_1^c = \emptyset,\ I_2 = \{1, 2\},\ I_2^c = \{3, 4\},$$

$$I_3 = \{1, 3\},\ I_3^c = \{2, 4\}.$$

The corresponding extremal correlation matrices are the following:

$$R_{I_1} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

$$R_{I_2} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix},$$

$$R_{I_3} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix},$$

and we have the following convex decomposition of the given correlation matrix:

$$R = 0.5R_{I_1} + 0.3R_{I_2} + 0.2R_{I_3}.$$

6. The quasi-extremal correlation matrices and their convex combinations

Let $X_1, \ldots, X_k$ be given variables having the same symmetrical distribution P_o.

DEFINITION 5. Let I^o be the initial index-set $\{1, 2, \ldots, k\}$. Say the sequence of index-sets $L = (L_1, \ldots, L_t)$ is *a t-partition* $(1 \leq t \leq k)$ *of the index-set* I^o, if the subsets $L_f = \{i_1^f, \ldots, i_{q_f}^f\}, f = 1, \ldots, t$ of I^o are non-overlapping and fulfill the following conditions:

$$i_1^{f_1} < i_1^{f_2}, \text{ if } f_1 < f_2,\ f_1 = 1, \ldots, t-1,\ f_2 = 2, \ldots, t. \qquad (9)$$

and

$$\bigcup_{f=1}^{t} L_f = I^o.$$

Hence the cardinalities q_f of the subsets L_f fulfill the following conditions:

$$\sum_{f=1}^{t} q_f = k.$$

The quasi-extremal correlation matrix, corresponding to a t-partition L, will be defined in the following way: all blocks, corresponding to subsets L_f, $f = 1, \ldots, t$, are extremal, but different blocks are independent. For exact formulation of the concept we need a definition of a more complicated partition of the initial index-set I^o (see Tiit, Tammet, 1994):

DEFINITION 6. Let L be a t-partition of an index-set I^o. Say (L, J) *is a t-double partition of the index-set* I^o, if for every subset L_f there is defined a two-partition (J_f, J_f^c) in such way that the subsets $J_f = (j_1^f, \ldots, j_{q_f}^f)$ and $J_f^c = (l_1^f, \ldots, l_s^f)$ of L_f are non-overlapping and fulfill the following conditions:

$$j_1^f = i_1^f, \quad J_f \cup J_f^c = L_f, f = 1, \ldots, t. \tag{10}$$

It is easy to see that the conditions (10) generalize the conditions (4) for the case of t-partition of the initial subset I_0.

DEFINITION 7. Let (L, J) be a t-double partition of the given index-set I^o. The correlation matrix $R^{(L,J)} = (r_{ij}^{(L,J)})$ is said to be *quasi-extremal*, if it is defined in the following way:

$$r_{ij}^{(L,J)} = \begin{cases} 0, \text{ if } i \in L_{f_1} \text{ and } j \in L_{f_2}, f_1 \neq f_2, \\ 1, \text{ if } i \in J_f \text{ and } j \in J_f \\ \quad \text{or } i \in J_f^c \text{ and } j \in J_f^c, f = 1, \ldots, t, \\ -1, \text{ if } i \in J_f \text{ and } j \in J_f^c \\ \quad \text{or } i \in J_f^c \text{ and } j \in J_f, f = 1, \ldots, t. \end{cases} \tag{11}$$

The defining formula (11) is an immediate generalization of the defining formula (5) for the case, when the correlation matrix consists of t independent blocks.

We shall use the symbol $\mathbf{R}^{(\mathbf{L},\mathbf{J})}(k)$ for denoting the set of all quasi-extremal correlation matrices of order k. If the order k is fixed, we shall write simply $\mathbf{R}^{(\mathbf{L},\mathbf{J})}$.

DEFINITION 8. Let R be a given correlation matrix of order k. If there exist such coefficients w_g, $g = 1, \ldots, h$, fulfilling the conditions (3), that the equation (2) holds, where $R_{I_g} \in \mathbf{R}^{(\mathbf{L},\mathbf{J})}(k)$, then we say that *for the correlation matrix R exists a convex decomposition by quasi-extremal correlation matrices.*

If the decomposition (2) exists, the decomposition with number of terms $h \leq k(k-1)/2 + 1$ will exist as well.

Between the sets $\mathbf{R}$ and $\mathbf{R}^{(\mathbf{L},\mathbf{J})}$ there exists evident inclusion $\mathbf{R} \subset \mathbf{R}^{(\mathbf{L},\mathbf{J})}$. Hence from the existence of a decomposition of given correlation matrix R by the extremal correlation matrices follows that the decomposition of the same correlation matrix R by the quasi-extremal correlation matrices exists as well. In the case, when the initial marginal distributions are equal and symmetrical (as it was supposed in this article), the opposite inclusion is true as well (see Tiit, 1986).

DEFINITION 9. Let L be a t-partition of the given index-set I^o. The correlation matrix $R^L = (r_{ij}^L)$ is said to be *quasi-maximal*, if it is defined in the following way:

$$r_{ij}^L = \begin{cases} 0, \text{ if } i \in L_{f_1} \text{ and } j \in L_{f_2}, f_1 \neq f_2, \\ 1, \text{ if } i \in L_f \text{ and } j \in L_f, \ f_1, f_2, f = 1, \ldots, t. \end{cases} \tag{12}$$

Formula (12) is a special case of the formula (11), when all 'negative' subsets J_f^c of parts L_f are empty, $f = 1, \ldots, t$.

We shall use the symbol $\mathbf{R^L}(k) = \mathbf{R^L}$ for denotation of the set of all quasi-maximal correlation matrices of order k.

It is easy to see, that quasi-maximal matrices form a subset of all quasi-extremal matrices.

In the case when the initial correlation matrix includes many elements close to zero, i.d. there are quite weak dependencies between several variables, it is reasonable to use the quasi-extremal correlation matrices instead of extremal correlation matrices in the convex decomposition of the given correlation matrix. In the case when the initial correlation matrix R has all positive elements it is usually useful to look for the decomposition of it in the set $\mathbf{R^L}$.

7. Computational procedure for decomposition of correlation matrix.

Let R be a given correlation matrix of arbitrary order k and let $R_1, \ldots, R_h$ be the given matrices (either from the set $\mathbf{R}$, $\mathbf{R}^L$ or $\mathbf{R^{LJ}}$) to be used in the convex decomposition. Then the decomposition problem can be solved as a solution of a system of linear equations

$$r_{ij} = \sum_{g=1}^{h} w_g r_{ij}^g, \tag{13}$$

where $i = 1, \ldots, k-1$, $j = i+1, \ldots, k$, by the additional conditions (3). Here the number h of unknowns w_g is, in general, much more than the number of equations, which equals to $k(k-1)/2+1$. The only exceptions form the cases $k \leq 3$, when the equation $k(k-1)/2+1 = 2^{k-1}$ holds.

Due to the following additional problems the solution of the system (3, 13) is quite complicated and labour-consuming.

1) The number of all quasi-extremal distributions is huge, and as it follows, the completing of equation system is rather complicated: for all used quasi-extremal correlation matrices $R^{(L,J)}$ all values of correlation coefficients should be calculated.

2) It is unknown beforehand if the exact solution exists at all.

But using the standard regression procedures, where the elements of the given correlation matrix are taken as the values (measurements) of the dependent variable and the elements of extremal or quasi-extremal correlation matrices as explanatory variables, we can get in many cases quite a satisfactory approximation to the exact decomposition.

8. Example 2. Comparison of decompositions in the case of the data of Tabachnic and Fidell

Here will be regarded the 'Small sample hypothetical data', introduced by Tabachnic and Fidell. The initial correlation matrix is the following (see Tabachnick, Fidell, 1989, pp 607 – 633):

$$\begin{pmatrix} COST & X_1 & 1 & -0.953 & -0.055 & -0.130 \\ LIFT & X_2 & -0.953 & 1 & -0.091 & -0.036 \\ DEPTH & X_3 & -0.055 & -0.091 & 1 & 0.990 \\ POWDER & X_4 & -0.130 & -0.036 & 0.990 & 1 \end{pmatrix}.$$

Using the traditional procedure of factor analysis the following two eigenvalues $\lambda_1 = 2.00$, $\lambda_2 = 1.91$ were found. The corresponding eigenvectors are:

$$h_1 = (-0.283, 0.177, 0.658, 0.675)',$$

$$h_2 = (0.651, -0.685, 0.252, 0.207).'$$

As it follows, the two-factor solution should be satisfactory, and these factors are:

$$F_1 = (-0.400, 0.251, 0.932, 0.956),'$$

$$F_2 = (0.900, -0.947, 0.348, 0.286).'$$

The first factor describes 50 % and the set of two factors 97.75 % of the diagonal elements of the correlation matrix, eg of the common variability of the initial data.

The sum of residual correlations

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(r_{ij} - \sum_{g=1}^{h} f_{ig}f_{jg})^2,$$

where for the factor loadings the denotation f_{ig} is used, equals to 0.0028 and the largest residual correlation has the absolute value equal to 0.005. From here it followos, that the two-factor solution is sufficient, e.g. the zero hypothesis that the residuals equal to zero, can be accepted for very large samples, having the size of order 100 000 measurements. For reasonable interpretation of the factor structure received some rotation procedure is needed (see Tabachnick, Fidell, 1989, pp 628 – 633).

Now let us use the convex-extremal decomposition for the same correlation matrix. At first we find the approximation by one quasi-extremal matrix. It is easy to see, that the matrix R_1, defined by the following (L, J)-partition: $L = (\{1,2\},\{3,4\})$, where $J_1 = \{1\}, J_1^c = \{2\}$ and $J_2 = \{3,4\}, J_2^c = \emptyset$, gives the best approximation for the initial matrix R. Let us write the matrix R_1 down by its elements:

$$R_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

For getting the first approximation due to the condition (3) no parameters should be estimated: the only parameter w_1 equals to one. From here it follows, that the first approximation is

$$R = R_1,$$

and the quality of approximation can be characterized by the sum S of the squares of residual correlation coefficients:

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(r_{i,j} - \sum_{g=1}^{h} w_g r_{i,j}^{L_g})^2. \qquad (14)$$

In this case the sum (see formula (14)) S equals to 0.08 and the largest residual has the absolute value 0.130. It means, that for the samples having size less than $n = 250$ it is possible to accept the zero hypothesis that the residuals equal to zeros.

The interpretation of the model is the following: the variables COST and LIFT are linearly dependent having the negative regression coefficient: when one of the values increases, the second decreases and vice versa. The last two variables DEPTH and POWDER are linearly dependent having positive regression coefficient. The both sets of variables are independent from each other.

Using the first approximation we must assume, that the structure of the initial data-set is entirely described by the quasi-extremal correlation matrix R_1.

For getting better approximation the second term must be added to the decomposition. This is the extremal correlation matrix R_2, corresponding to the partition $(\{1,2\},\{3,4\})$:

$$R_2 = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{pmatrix}.$$

To estimate the parameters the traditional regression analysis procedure was used, where $Y := r_{ij}$ and $X_g := r_{ij}^{L_g}$, $g = 1,\ldots,h$. The model without constant term will be used, and the estimated regression coefficients b_g will be taken as weights w_g. For warranting the satisfaction of the condition (3) the following additional transformation should be done:$Y := Y - X_h; X_g := X_g - X_h, h = 1,\ldots,h-1$. Then $w_h = 1 - \sum_{g=1}^{h-1} w_g$. The solution is acceptable only if if the weights (regression coefficients) are nonnegative. That means, from all possible solutions only the acceptable ones will be regarded.

The second approximation of the initial correlation matrix is the following:

$$R = 0.933R_1 + 0.067R_2,$$

and the intepretation is: 93.3% of the population has the dependence structure, characterized by the correlation matrix R_1 and 6.7 % has the dependence structure, characterized by the correlation matrix R_2. In this case the sum (14) S of the squares of deviations equals to 0.027 and the zero hypothesis about the zero residuals can be accepted in the case when the population size is not more than 500.

9. Discussion

A. The preferences of the convex-extremal decomposition are the following:

I. The set of elements of the extremal/quasi-extremal correlation matrices is standard and does not depend on the concrete sample.

II. The number h of parameters (weights) to estimate is much less than in the case of factor decomposition, when $h \cdot k$ factor loadings must be estimated, where h is the number of factors.

III. The mixture of distributions can be interpreted as a mixture of populations and hence it is quite easy to understand.

IV. No rotation procedure is needed for improving the solution.

In the second example the both decompositions gave quite different results. As the assumptions of the Theorem 5 were not fulfilled, the equivalent decompositions did not exist. It is remarkable, that in this case *one quasi-extremal matrix* gave almost as good description as *two factors*. For the factor decomposition 8 parameters (the factor loadings) had to be estimated, but in the case of the convex-extremal decomposition the only weight w_1 was equal to one, so *no estimation was needed at all.* The reason of this feature is, that the quasi-extremal correlation matrices are able to describe more complex dependence structures than the factors (and the extremal correlation matrices, which are in some sense equivalent to each other).

From here it follows, that it would be reasonable to use the convex-extremal decomposition especially for the case of small samples.

B. The shortcomings of the convex-extremal decomposition compared with the factor decomposition are the following:

I. For all correlation matrices the exact convex-extremal decomposition (2) - (4) does not exist at all, but the decomposition (1), forming the basis for the factor decomposition, always exists.

II. The convex decomposition is not unique in the sense that there might be different sets of extremal or quasi-extremal correlation matrices $\{R_{I_1}, \ldots, R_{I_j}\}$, satisfying the conditions (2) - (4) either exactly, either approximately.

In fact, the same situation occurs in factor decomposition in the case of equal or close eigenvalues, but also in using different numbers of factors, different factors extraction or different rotation procedures.

III. For factor analysis there exists standard software and long tradition of usage.

In the solution of computational problems of convex-extremal decomposition there is a possibility to use standard statistical software (multiple and step-wise regression procedures) as it was men-

tioned above.

10. Example 3. The analysis of changes in the anthropological structure of Estonian schoolchildren.

To analyse the development of Estonian children's body structure in age, some measurements were carried out among schoolchildren aged 6 to 18, both boys and girls. In each age and sex group there were taken about 150 to 200 children. The sample of schools was representative of Estonia (rural/urban, different regions). Only Estonian speaking population was considered. From every schoolchild about 60 different measurements were taken, but for more detailed investigation only 12 in some sense most informative of them were chosen.

The measurements used in this study were the following (in brackets is the number of variable): *weight* (1), *height* (2), *cervical height* (3), *foot length* (4), *upper limb length* (5), *lower limb length* (6), *chest circumference* (7), *pelvis circumference* (8), *biacromial breadth* (9), *chest breadth* (10), *chest depth* (11), *pelvis breadth* (12). On the basis of these data 26 correlation matrices of order 12 were calculated, one for each age-sex group. All of them were quite similar, having only quite large positive correlation coefficients (on average the range of them was 0.25 - 0.95), practically all correlations were significant. Using component or factor analysis in all subpopulations several principal components/factors can be found, but there is no effective procedure for establishing and modelling *the changes* in the factor structures of children of different age groups.

Then the convex decomposition of correlation matrices by quasi-extremal ones was used. The assumptions used in the study were the following: All anthropological measurements have the distributions, quite close to the normal distributions, where the deviances from the normality are almost the same. character - small positive skewness, very small excess (see Kaarma, 1981). Hence the assumption about the equal distributions (up to the linear transformation), see (3) is valid. As all correlations are positive, it is reasonable to use the decomposition by *quasi-maximal* distributions. In this case the assumption about symmetry of all marginals is not necessary. The main advantage of the extremal decomposition behind the principal decomposition is the existence of *the same standard base elements* for all correlation matrices to be considered – these are the quasi-maximal correlation matrices, consisting of

ones and zeros only.

In this case the construction of the model of change can be carried out by analysing the change of coefficients in the decompositions.

For getting the comparable decompositions, several quasi-positive correlation matrices, which were common in decompositions of different age-groups, were chosen.

I. As the first step, the decomposition by three matrices R_1, consisting of all 1's, $R_0 = I$ (unit matrix) and R_2, corresponding to the following partition:

$$L = \{\{1,7,8\},\{2,3,5,6\},\{9\},\{10\},\{11\},\{12\}\}.$$

To this partition the following correlation matrix corresponds:

$$R_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

was made. For example, in the case of the six-year old boys we received the following decomposition:

$$R = 0.158R_0 + 0.613R_1 + 0.229R_2. \tag{15}$$

One possible interpretation of the model (15) is the following. The population of 6-year-old boys can be divided by the structure of their body into three groups. The first group corresponds to R_1 – all body measurements of the boys of this group are linearly dependent (proportional). Hence the body structure of all boys of this group is quite similar, but their sizes can be different. From the decomposition formula (15) it follows, that about 61% of all boys belong to this group.

The second group, corresponding to R_2, can be characterized through two sets of proportional measurements – to one of them belong the measurements, characterizing the tallness of boys - height, cervival height, upper and lower limb lengths. The second set consists of measurements, characterizing the thickness of boys - weight, chest and pelvis circumferences. It is typical, that the variables, belonging to different sets, are uncorrelated (independent). These measurements are uncorrelated with all other measurements (*foot length*, biacromial breadth etc, as well. Hence into this group belong the boys, who can be either thick either thin, and are characterized as pycnic and leptosomous anthropometrical types. The rate of 6-year old boys, belonging to these types, is about 23% .

The third group corresponds to the unit correlation matrix, that means, all body measurements can be considered as uncorrelated. In fact, we can assume that the correlations between body measurements are rather weak and the higher correlations might be situated in random cells of the correlation matrix. As it follows from the formula (15), the percentage of such 6-year old boys is about 16%.

In the similar way the correlative matrices characterizing the body structure of boys and girls of all age groups were found. As a result, two 3-dimensional time series – one $T(B)$ for boys and the second, $T(G)$ for girls,

$$T(B) = (t_i^0(B), t_i^1(B), t_1^2(B)),\ T(G) = (t_i^0(G), t_i^1(G), t_i^2(G))$$

$$i = 6, \ldots, 18,$$

were received. Here the index i characterizes the age, and the following general condition,

$$\sum_{j=1}^{2} t_i^j(A) = 1,\ i = 6, \ldots, 18,\ A = B, G$$

is fulfilled.

These time series are illustrated by the figures 1 and 2, where the so-called *component charts* of the series are given.

We see the following trends in the coefficients:

1) The rate of the *full dependence set*, characterized by the correlation matrix R_1, eg the *entirely proportional body types*, decreases for both, boys and girls, about 1.2 to 1.5 per cent by year, the change is statistically significant, $p < 0.05$.

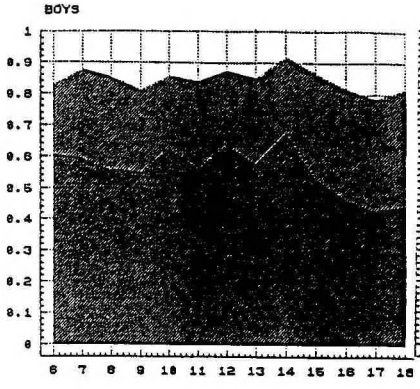

Figure 1.

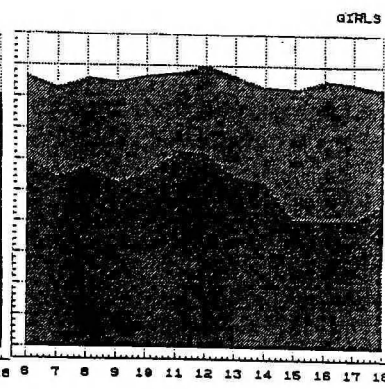

Figure 2.

Changes of weights (w_g) of different extremal correlation matrices, i.e. shares of different body structures in age from 6 till 18 years.

w_1 – entirely proportional measurements,

w_2 – classical anthropological types,

$w_0 = 1 - (w_1 + w_2)$ – non-proportional (independent) measurements.

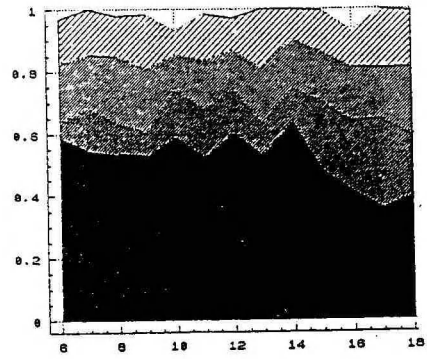

Figure 3.

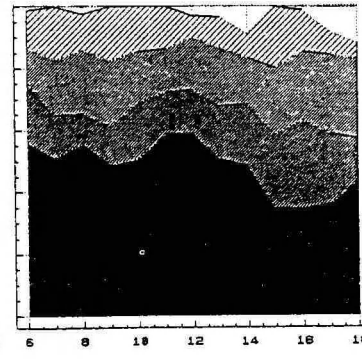

Figure 4.

Changes of weights (w_g) of different extremal correlation matrices.

w_1 – entirely proportional measurements,

w_2 – classical anthropological types,

w_3 – classical anthropological types (all measurements included),

w_4 – weakly dependent measurements,

$w_0 = 1 - (w_1 + w_2 + w_3 + w_4)$ – non-proportional (independent) measurements

2) The other characteristic subpopulation, described by the correlation matrix R_2, eg the children, *belonging to different anthropometrical types*, increases, in average about one per cent by year, wherby the change is strongly significant ($p < 0.01$).

II. More detailed characteristic of anthropometrical types can be received using more quasi-maximal correlation matrices. On figures 3 and 4 the component charts of the decompositions by 5 matrices are given. The matrix R_3, corresponding to the partition

$$L_3 = \{\{1,7,8,10,11\},\{2,3,4,5,6\},\{9,12\}\}$$

and matrix R_4, corresponding to one maximally correlated subset $\{1,2,3,12\}$ and hence to the partition

$$L_4 = \{\{1,2,3,12\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\},\{11\}\}$$

were added to the matrices, used in the partition (15).

We see, that the same general tendencies as in the case of decomposition (15) are obvious, but some new features occur:

3) The group, characterizing the *anthropometrical types* (i.d., corresponding to R_2 in the decomposition (15)) is divided into two subpopulations, corresponding to matrices R_2 and R_3. In the last one all measurements participate in dependencies: they are divided into three groups:

a) the measurements of length (tallness) – *height, cervival height, foot and limbs' lengths*;

b) the measurements of thickness (stockyness) - *weight, chest and pelvis circumference, chest breadth and depth*;

c) the measurements of breadth: *biacromial breadth and pelvis breadth.*

4) Part of the subpopulation, characterized by the matrix R_0 (independendent measurements) can be characterized by somewhat more rich dependency structure, described by matrix R_4, indicating the correlations between *weight, height, cervival height* and *pelvis breadth.* The children of this group cannot be considered as representatives of some typical anthropometrical types, but their body weight is correlated with length and pelvis breadth, hence the main measurements are proportional.

Here again we see that the rate of anthropometrical types increases with age for boys and girls, but the rate of simply proportional bodies – decreases. Also the fact, that in the case of older

teenagers more measurements are involved to the typical dependence structures seems to be quite reasonable conclusion 2 proved by the given statistical procedure.

References.

Frechet, M. (1951) Sur les tableux de correlatioon dont les marges sont donnees. Ann. Univ. Lyon. Sect. A, **14**, 53 - 77.

Hoeffding, W. (1940) Masstabinvariante Korrelations theorie. Schr. Math. Inst. u. Inst. Angew. Math. Univ. Berlin, **5**, 179 - 233.

Kaarma,H. (1981) Multivariate statistical investigation of the system of anthropometrical variables of women. Tallinn, 'Valgus' (In Russian).

Kotz,S., Tiit, E.-M. (1992) Bounds in multivariate dependence. Acta et Commentationes Universitatis Tartuensis, **942**, 35 - 45.

Tabachnick, B.G., Fidell, L.S. (1989) Using Multivariate Statistics. Second Edition, N York, Harper Collins Publishers.

Tiit, E.-M. (1984) Definition of random vectors with given marginal distributions and given correlation matrix. Acta et Commentationes Universitatis Tartuensis, **685**, 21 - 36.

Tiit, E.-M. (1986) Random vectors with given arbitrary marginals and given correlation matrix. Acta et Commentationes Universitatis Tartuensis, **733**, 14 - 38.

Tiit, E.-M. (1992) Extremal multivariate distributions having given discrete marginals. Acta et Commentationes Universitatis Tartuensis, 942, p. 94 - 113.

Tiit, E.-M., Tammet, S. (1994) Using index-vectors and partitions in multivariate analysis. This issue, p. 55 - 75.

Received November, 1993.

Korrelatsioonimaatriks kumer-ekstremaalne lahutus. Rakendus antropomeetrilistel andmetel

Ene-Margit Tiit, Maie Thetloff

Kokkuvõte

Käesolevas artiklis vaadeldakse ekstremaalseid ja kvaasiekstremaalseid jaotusi ning vastavaid korrelatsioonimaatrikseid. Viimaste elementideks on kas ühed ja miinus ühed, kvaasiekstremaalsete korrelatsioonimaatriksite elementideks võivad olla ka nullid. Artiklis võetakse kasutusele antud korrelatsioonimaatriksi kumer lahutus ekstremaalsete (kvaasiekstremaalsete) korrelatsioonimaatriksite kaudu. Jaotuste klassis vastab sellele ekstremaaljaotuste segu, mis on antud mitmemõõtmelise jaotuse lähendiks.

Korrelatsioonimaatriksi kumerat lahutust võrreldakse faktoranalüüsiga ning leitakse tingimused, millal mõlemad lahutused ühtivad. Tuuakse näiteid nii ühtiva kui erineva lahutuse kohta ja analüüsitakse mõlema meetodi tugevaid ning nõrku külgi. Tõdetakse, et kumer-ekstremaalne jaotus sobib hästi seoste struktuuride võrdlemiseks erinevates andmestikes.

Rakenduslikus näites vaadeldakse eesti kooliõpilaste (vanuses 6 - 18 aastat) kehaehitustüüpide vahekordade muutusi sõltuvalt vanusest. Leitakse statistiliselt olulised nihked – vanuse suurenedes suureneb klassikalistesse kehaehituse tüüpidesse sobivate noorukite hulk lastega võrreldes.

Acta et Commentationes Universitatis Tartuensis, **968** ,1994, pp. 99-104

Cornish–Fisher Expansion for the Ratio of Two Sample Means

Imbi Traat

Summary. The paper gives the explicit expression of the Cornish-Fisher expansion with the accuracy $o(n^{-1/2})$ for the quantile of the ratio of two sample means in the case of i.i.d. observations.

Keywords: *Edgeworth expansion, Cornish–Fisher expansion, ratio of sample means, approximation accuracy*

1. Introduction

The ratio of sample means is a frequently used statistic in many application areas, e.g. in Survival Analysis, Survey Sampling etc. Its distribution is usually unknown and therefore several approximations are looked for. Much attention has been paid to the Edgeworth expansion of its distribution. Validity of the expansion has been proved under more and more general and weaker conditions, see e.g. Bai and Rao (1992). In applications, instead of distribution or density function, very often the corresponding quantile is needed, e.g. for constructing confidence intervals at a given level. It is possible to approximate the quantile of an unknown distribution by the Cornish–Fisher type expansion first introduced by Cornish and Fisher (1937). A simple derivation technique of a general Cornish–Fisher expansion is given by Lee and Lee (1992). On the basis of an example the authors also show, that for certain distributions, quantile approximation gives even better approximation of the probability than Edgeworth expansion does. In the present paper we give the formal Cornish–Fisher expansion up to the order $o(n^{-1/2})$ for the quantile of the ratio of two sample means.

Let (x_i, y_i), $i = 1, ..., n$ be i.i.d. observations of the random vector (x, y). Denote

$$Ex = \mu, \ Ey = \eta, \ (\eta \neq 0)$$

$$\eta^{-4} E(\eta x - \mu y)^2 = \sigma^2.$$

We consider the standardized quantity

$$S_n = \sqrt{n}\sigma^{-1}(\frac{\bar{x}}{\bar{y}} - \frac{\mu}{\eta}). \tag{1.1}$$

Bai and Rao (1992) have shown that if the characteristic function $v(t)$ of $\eta x - \mu y$ satisfies the Cramér c-condition

$$\overline{\lim}_{|t|\to\infty} |v(t)| < 1 \tag{1.2}$$

and x, y have finite m-th ($m \geq 3$) moment, then the distribution function F_n of S_n is approximated uniformly by the Edgeworth expansion U_{nm} with the accuracy $o(n^{-(m-2)/2})$. Cramér condition for $\eta x - \mu y$ allows one of the variables x, y to be discrete.

2. Cumulants of the ratio of two sample means

To find either Edgeworth expansion or Cornish–Fisher expansion for a statistic the most difficult task usually is to find the cumulants of that statistic. They are often found as power series of n, where the coefficients are some functions of the cumulants of the parent population.

Bai and Rao (1992) give the explicite Edgeworth expansion for S_n using the cumulants of $\sigma^{-1}\eta^{-2}(\eta x - \mu y)$ and $\eta^{-1}(y - \eta)$. To express the cumulants of S_n through the cumulants of x and y, we use ordinary Taylor expansion method. The Taylor expansion of $\bar{x}/\bar{y}$ around μ/η is

$$\frac{\bar{x}}{\bar{y}} = \frac{\mu}{\eta} + \frac{1}{\eta}(\bar{x} - \mu) - \frac{\mu}{\eta^2}(\bar{y} - \eta)$$

$$+\frac{\mu}{\eta^3}(\bar{y} - \eta)^2 - \frac{1}{\eta^2}(\bar{x} - \mu)(\bar{y} - \eta) + \dots .$$

Finding necessary expectations we get the following first cumulants of S_n:

$$ES_n = \lambda_1 + O(n^{-3/2}),$$

$$E(S_n - ES_n)^2 = 1 + O(n^{-1}),$$
$$E(S_n - ES_n)^3 = \lambda_3 + O(n^{-3/2}),$$

where

$$\lambda_1 = n^{-1/2}\sigma^{-1}\eta^{-3}(\mu m_{02} - \eta m_{11}), \tag{2.1}$$

$$\lambda_3 = n^{-1/2}\sigma^{-3}\eta^{-6}(\eta^3 m_{30} - 3\mu\eta^2 m_{21} + 3\mu^2\eta m_{12} - \mu^3 m_{03}), \tag{2.2}$$

and

$$m_{ij} = E(x-\mu)^i(y-\eta)^j.$$

The Edgeworth expansion U_{n3} for S_n has then the following explicite form:

$$U_{n3} = \Phi(x) - (\lambda_1 + \lambda_3(x^2-1)/6)\dot{\phi}(x) + o(n^{-1/2}), \tag{2.3}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are standard normal distribution function and density function respectively.

3. Cornish–Fisher expansion for the quantile of S_n

We use here the method of Lee and Lee (1992) to derive the expansion of the quantile of S_n including the term $n^{-1/2}$. The general idea stands in looking for the relationship between arguments z and ξ so that

$$\Phi(\xi) = F(z). \tag{3.1}$$

The distributions Φ and F are expected to be close to each other so that

$$z = \xi + \delta.$$

Then the Taylor expansion of F

$$F(\xi+\delta) = \sum_{k=0}^{\infty} \frac{1}{k!} D^k F(\xi)\delta^k = \exp(\delta D)F(\xi), \tag{3.2}$$

is used, where $D^k F(\xi)$ means $(d^k/dt^k)F(t)|_{t=\xi}$. Distribution function $F(\xi)$ on the right–hand side is substituted by its Edgeworth expansion

$$F(\xi) = \exp(\sum_{k=1}^{\infty} \lambda_k(-D)^k/k!)\Phi(\xi), \tag{3.3}$$

where the cumulants of the distribution F are of following orders:

$$\lambda_1 = \lambda_2 = 0, \lambda_k = O(n^{1-k/2}), k \geq 3\,.$$

The quantity δ is expected to have the form

$$\delta = \delta_1 + \delta_2 + \delta_3 + ... \tag{3.4}$$

with $\delta_k = O(n^{-k/2})$. From (3.1)-(3.4) Lee and Lee get the equation where F is eliminated. Expanding then the exponent and using the partition of an integer they write down the general expression for δ_k.

In our case the cumulants of S_n have different orders of n, all being power series of $n^{-1/2}$. For such a case the explicite Cornish–Fisher expansion is given by Withers (1984). Coefficients in these power series certainly depend on considered statistic and they need to be found for special cases. To present the Cornish–Fisher expansion for our special case we use the idea of Lee and Lee (1992). Using only the terms of order $O(n^{-1/2})$, we obtain from (3.1)-(3.4)

$$\Phi(\xi) = \exp(\delta_1 D - \lambda_1 D - \lambda_3 D^3/6...)\Phi(\xi).$$

After expanding we have

$$\Phi(\xi) = (1 + \delta_1 D - \lambda_1 D - \lambda_3 D^3/6...)\Phi(\xi).$$

Using Hermite polynomials, defined by

$$D\Phi(\xi) = \phi(\xi),$$

$$D^3\Phi(\xi) = H_2(\xi)\phi(\xi) = (\xi^2 - 1)\phi(\xi),$$

where $\phi(\xi)$ is the density function of N(0,1), we have

$$\delta_1 = \lambda_1 - \lambda_3(\xi^2 - 1)/6.$$

Equation (3.1) being equal to α means that there exists a relationship between α-quantile of F and α-quantile ξ_α of standard normal distribution. For the α-quantile of S_n we have obtained the following Cornish–Fisher expansion up to the order $o(n^{-1/2})$:

$$z_\alpha = \xi_\alpha + \lambda_1 + \lambda_3(\xi_\alpha^2 - 1)/6, \tag{3.5}$$

where λ_1, λ_3 are determined by (2.1)-(2.2).

To say something about the accuracy of this quantile expansion, notice that

$$P(S_n < z_\alpha) = P(S_n - \lambda_1 - \lambda_3(\xi_\alpha^2 - 1)/6 < \xi_\alpha).$$

Since the random variable

$$S_n - \lambda_1 - \lambda_3(\xi_\alpha^2 - 1)/6$$

has the same cumulants as S_n, except the first cumulant (mean), which is equal to

$$-\lambda_3(\xi_\alpha^2 - 1)/6,$$

the $n^{-1/2}$ -term in its Edgeworth expansion vanishes and we get

$$P(S_n < z_\alpha) = \Phi(\xi_\alpha) + o(n^{-1/2}) = \alpha + o(n^{-1/2}).$$

So we have verified the following:

Theorem. Suppose that random variables x, y have finite third moments ($E(y) \neq 0$) and the characteristic function $v(t)$ of $\eta x - \mu y$ satisfies the Cramér c-condition (1.2). Then the standardized ratio of sample means

$$S_n = \sqrt{n}\sigma^{-1}(\frac{\bar{x}}{\bar{y}} - \frac{\mu}{\eta})$$

has the α–quantile

$$z_\alpha = \xi_\alpha + \lambda_1 + \lambda_3(\xi_\alpha^2 - 1)/6$$

with the accuracy

$$P(S_n < z_\alpha) = \alpha + o(n^{-1/2}).$$

References

Bai, Z. D., Rao, C. R. (1992). A note on Edgeworth expansion for ratio of sample means. Sankhyā, Series A, **54**, 309–322.

Cornish, E. A., Fisher, R. A. (1937). Moments and cumulants in the specification of distributions. Rev. Internat. Statist. Inst. **5**, 307–320.

Lee, Y. S., Lee, M. C. (1992). On the derivation and computation of the Cornish–Fisher expansion. Austral. J. Statist., **34** (3), 443–450.

Withers, C. S. (1984). Asymptotic expansions for distributions and quantiles with power series cumulants. J. R. Statist. Soc. B, **46**, 3, 389–396.

Received January, 1994

Cornish–Fisheri reaksarendus valimikeskmiste suhte jaoks

Imbi Traat

Kokkuvõte

Vaadeldakse juhuslikku vektorit (x, y), mis omab lõplikke kolmandaid momente ja rahuldab Craméri c–tingimust (1.2). Vektorist on antud n–elemendiline valim ja vastav valimikeskmiste normeeritud suhe S_n (1.1). Statistikut S_n kasutatakse mitmetes rakendustes, näiteks elukestuse analüüsis, valikuuringus jm. Praktilistes ülesannetes eeldatakse tavaliselt, et S_n on normaaljaotusega ja usalduspiiride leidmisel kasutatakse tema α–kvantiilina normaaljaotuse α–kvantiili. Üldjuhul, eriti aga väikeste valimite korral, ei ole S_n normaaljaotusega, kusjuures pole ka teada tema tegelik jaotus. Artiklis esitatakse statistiku S_n ligikaudne α–kvantiil z_α – st. Cornish-Fisheri reaksarendus (3.5). Ligikaudse kvantiili viga koondub protsessis $n \to \infty$ nulliks kiiremini kui $n^{-1/2}$, st. $P(S_n < z_\alpha) = \alpha + o(n^{-1/2})$.