Chun-houh Chen

(1)

A THEOREM ABOUT CONCOR

Joseph B Kruskal
Bell Laboratories
Murray Hill, N.J.

Suppose that P is the (pattern) matrix of $+1$'s and -1's which indicates the division of the n objects into two blocks (in the usual CONCOR fashion). Suppose that A is the current matrix of correlations, and let dev(A) indicate the the largest absolute difference between an entry of A and the corresponding entry of P. We prove that if dev(A) is small enough, then further iteration will necessarily lead to the correlation matrix converging to P. Specifically, suppose that

n is the size of the matrices,

k and $n-k$ are the sizes of the two blocks,

$$e = 8\,\frac{k}{n}\left(1 - \frac{k}{n}\right),$$

$$\mathrm{dev}(A) < \frac{-1 + \sqrt{1 + 4e^2}}{2e}.$$

Then the correlation matrix must converge to P.

21 September 1977

For example, suppose that the $n=20$ objects are split into two blocks of size $k=5$ and $n-k=15$. Then we have

$$e = 8(1/4)(3/4) = 3/2,$$

$$\text{dev}(A) < \frac{-1 + \sqrt{1+9}}{3} = \frac{-1 + 3.16}{3} = .72 \ .$$

Thus if all the correlations corresponding to $p_{ij}=+1$ are greater than $.28$ and all the correlations corresponding to $p_{ij} = -1$ are smaller than $-.28$, then the theorem applies.

DETAILED STATEMENT: Matrices are n by n. A indicates a matrix with entries a_{ij}, and we assume that $-1 \leqq a_{ij} \leqq +1$. Let I_1 be a subset of $\{1, 2, \ldots, n\}$ which contains k members, and let I_2 be its complement. $P = \| p_{ij} \|$ is the matrix such that

$$p_{ij} = \begin{cases} +1 & \text{if } i \text{ and } j \text{ both belong to } I_1 \\ +1 & \text{if } i \text{ and } j \text{ both belong to } I_2 \\ -1 & \text{otherwise} \end{cases}.$$

Define $\operatorname{dev}(A)$ by

$$\operatorname{dev}(A) = \max_{i,j} |a_{ij} - p_{ij}| .$$

Define A^* to be the matrix such that

a^*_{ij} = the correlation between columns i and j of A.

We also write $A^{(0)} = A$, $A^{(1)} = A^*$, and

$A^{(n)} = (A^{(n-1)})^* = A^{*\cdots*}$ with n stars.

Define a function $\rho(\delta)$ for $0 \leqq \delta \leqq 1$ by

$$\rho(\delta) \equiv_{def} \frac{1}{8 \frac{k}{n}(1-\frac{k}{n})} \frac{\delta}{1-\delta^2} .$$

THEOREM: Let $\delta = \operatorname{dev}(A)$, and suppose that $\delta < 1$. Then

$$\operatorname{dev}(A^*) \leqq \rho(\delta) \cdot \delta \equiv \frac{1}{8 \frac{k}{n}(1-\frac{k}{n})} \frac{\delta^2}{1-\delta^2} , \quad \square$$

We make use of the theorem through two corollaries. The second one supports the assertion given on the first page.

COROLLARY 1: Let $e = 8 \frac{k}{n}\left(1-\frac{k}{n}\right)$. If $\delta < \frac{-1+\sqrt{1+4e^2}}{2e}$, then $\frac{\delta}{1-\delta^2} < e$ and $\rho(\delta) < 1$, so

$$\mathrm{dev}(A^*) \leqq \mathrm{dev}(A).$$

If $\delta \neq 0$, the inequality is strict. $\square$

COROLLARY 2: Under the same hypothesis as above,

$$\mathrm{dev}(A^{(n)}) \leqq [\rho(\delta)]^n \, \mathrm{dev}(A)$$

and therefore $A^{(n)} \to P$ as $n \to \infty$. $\square$

PROOF OF THEOREM: We want to prove an upper bound on the numbers $|a^*_{ij} - p_{ij}|$. Noting that p_{ij} [redacted] always is ± 1, it is enough to prove an upper bound for the numbers $|p_{ij}(a^*_{ij} - p_{ij})| = |p_{ij} a^*_{ij} - 1|$. Thus it suffices to show that

$$|p_{ij} a^*_{ij} - 1| \leq \frac{1}{8\frac{k}{n}\left(1-\frac{k}{n}\right)} \quad \frac{\delta^2}{1-\delta^2}$$

for all i, j. This is trivial for $i=j$. Without real loss of generality we take $i=1$ and $j=2$.

Thus consider a^*_{12}. This is defined as the correlation between a_{i1} and a_{i2}. Let $\bar{a}_j = \frac{\sum_i a_{ij}}{n}$.

Then

$$\sum_i \left[(a_{i1} - \bar{a}_1) - p_{12}(a_{i2} - \bar{a}_2)\right]^2 \leq \sum_i \left[a_{i1} - p_{12} a_{i2}\right]^2 \leq n\,\delta^2.$$

The first inequality holds because the second moment of any set of numbers is smallest when taken around their mean (as compared with the moment around any other value).

We prove the second inequality by proving that for every i, $|a_{i1} - p_{12}a_{i2}| \leq \delta$. To see this, recall the definition of $dev(A)=\delta$, and note that each a_{ij} lies either in the interval $[-1, -1+\delta]$ or $[1-\delta, 1]$. Of course $p_{12}a_{i2}$ lies in one of these intervals also. It suffices to show that a_{i1} and $p_{12}a_{i2}$ belong in the same one of these intervals.

For every i' and j', $a_{i'j'}$ lies in the same interval as $p_{i'j'}$, so it suffices to show that p_{i1} and $p_{12}p_{i2}$ lie in the same interval. Recalling the definition of P, based on the complementary sets I_1 and I_2, it is easy to show by considering cases that if i and 1 belong to the same one of these sets then $p_{i1} = 1 = p_{12}p_{i2}$ and otherwise $p_{i1} = -1 = p_{12}p_{i2}$.

We expand the inequality to obtain the following:

$$\Sigma(a_{i1}-\bar{a}_1)^2 + \Sigma(a_{i2}-\bar{a}_2)^2 - 2\,p_{12}\sum_i (a_{i1}-\bar{a}_1)(a_{i2}-\bar{a}_2) \leq n\delta^2 .$$

Let $s_j = \sqrt{\frac{1}{n}\Sigma(a_{ij}-\bar{a}_j)^2}$. Then

$$n s_1^2 + n s_2^2 \leq 2\,p_{12}\,\Sigma(\,)(\,) + n\delta^2 .$$

Dividing by $2n s_1 s_2$,

$$\frac{1}{2}\left(\frac{s_1}{s_2} + \frac{s_2}{s_1}\right) \leq p_{12}\underbrace{\frac{\frac{1}{n}\Sigma(\,)(\,)}{s_1 s_2}}_{\overset{\shortparallel}{a^*_{12}}} + \frac{1}{2}\,\frac{\delta^2}{s_1 s_2}$$

Now

$$\frac{s_1}{s_2} + \frac{s_2}{s_1} \geq 2, \text{ so}$$

$$1 \leq p_{12}\,a^*_{12} + \frac{1}{2}\,\frac{\delta^2}{s_1 s_2}$$

$$p_{12}\,a^*_{12} \geq 1 - \frac{1}{2}\,\frac{\delta^2}{s_1 s_2} .$$

Now we wish to obtain a lower bound for s_j which we can substitute into this inequality. It is intuitively evident that if we pull back the a_{ij} near 1 to $1-\delta$ and the values near -1 to $-1+\delta$, the standard deviation based on these will be smaller than s_j. More formally, we define $\tilde{s}_j$ as follows:

$$\tilde{s}_j = \sqrt{\frac{1}{n} \sum \left\{ [p_{ij}(1-\delta)] - \tilde{p}_j \right\}^2}$$

where $\tilde{p}_j = \frac{1}{n} \sum_i [p_{ij}(1-\delta)] = \frac{2k-n}{n}(1-\delta)$.

A formal proof that $s_j \geq \tilde{s}_j$ can easily be obtained by noting that the projection mapping of the reals onto the closed interval $[-1+\delta, 1-\delta]$ is a contraction mapping and by using the well-known elementary identity

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{2n} \sum_i \sum_j (x_i - x_j)^2 .$$

Thus we have

$$s_j^2 \geq \tilde{s}_j^2 = \frac{\sum [p_{ij}(1-\delta)]^2}{n} - \tilde{p}_j^2$$

$$= \frac{n(1-\delta)^2}{n} - \left[\frac{2k-n}{n}(1-\delta)\right]^2$$

$$= (1-\delta)^2 \left[1 - \left(1 - \frac{2k}{n}\right)^2\right]$$

$$= 4(1-\delta)^2 \frac{k}{n}\left(1 - \frac{k}{n}\right).$$

Therefore

$$s_1 s_2 \geq 4(1-\delta)^2 \frac{k}{n}\left(1-\frac{k}{n}\right),$$

so

$$p_{12} a_{12}^* \geq 1 - \frac{1}{2} \frac{\delta^2}{4(1-\delta)^2 \frac{k}{n}\left(1-\frac{k}{n}\right)}$$

$$p_{12} a_{12}^* \geq 1 - \frac{1}{8 \frac{k}{n}\left(1-\frac{k}{n}\right)} \frac{\delta^2}{1-\delta^2}.$$

This completes the proof of the theorem. $\square$

PROOF OF COROLLARY 1: It is plain that

$e = 8 \frac{k}{n} (1 - \frac{k}{n})$ is positive. From this it is easy to show that $\frac{-1 + \sqrt{1+4e^2}}{2e} < 1$. Therefore the hypothesis shows $\delta < 1$, and $1-\delta^2$ is positive. Given this, it follows that the inequality $\frac{\delta}{1-\delta^2} < e$ is equivalent to

$$e\delta^2 + \delta - e < 0.$$

The two values of δ for which this quadratic function of δ is 0 are $\frac{-1 \pm \sqrt{1+4e^2}}{2e}$. One of these solutions is positive and one negative. The values of δ for which the quadratic function is negative lie between these two values, and hence include the interval from 0 to $\frac{-1 + \sqrt{1+4e^2}}{2e}$. Therefore the hypothesis of the corollary implies $\frac{\delta}{1-\delta^2} < e$ and hence $\rho(\delta) < 1$.

The theorem now yields that

$$\mathrm{dev}(A^*) \leq \rho(\delta) \cdot \mathrm{dev}(A) \leq \mathrm{dev}(A).$$

The latter inequality is strict if $\delta \neq 0$. □

PROOF OF COROLLARY 2: If $\delta \equiv \text{dev}(A) = 0$, then the result is trivial, so we assume $\delta \neq 0$.

The result is trivial for n=0, and follows from the theorem for n=1. We proceed by induction, and assume that the result is true for some value n. Now by the theorem,

$$\text{dev}\left(A^{(n+1)}\right) \leqq \rho(\delta_n) \cdot \delta_n$$

where

$$\delta_n = \text{dev}\left(A^{(n)}\right).$$

By the induction hypothesis, and because $\rho(\delta) < 1$,

$$\delta_n \leqq [\rho(\delta)]^n \delta \leqq \delta .$$

Since ρ is an increasing function

$$\rho(\delta_n) \leqq \rho(\delta)$$

Using these inequalities, we obtain

$$\text{dev}\left(A^{(n+1)}\right) \leqq \rho(\delta) \cdot [\rho(\delta)]^n \delta = [\rho(\delta)]^{n+1} \delta .$$

The fact that $A^{(n)} \to P$ is now trivial. $\square$