CONVERGENCE OF THE ITERATIVE SCALING PROCEDURE FOR NON-NEGATIVE MATRICES

OLIVER PRETZEL

The iterative scaling procedure (or iterative proportional fitting procedure) was first suggested in 1940 by Deming and Stephan [4]. It consists in modifying a non-negative matrix to achieve specified row and column sums by alternately multiplying the rows and columns to adjust their sums to the desired values. Since 1940 the procedure has been the subject of numerous papers both by statisticians and pure mathematicians. Fienberg [6] contains an admirable review of the papers up to 1970.

Sinkhorn [11] first stressed the idea of diagonal equivalence in 1964 and it is clear that by the nature of the process the iterates are diagonally equivalent to the original matrix. That is also true for the limit if no non-zero cell of the initial matrix tends to zero (an easy proof of this is included in the present paper for completeness). Suppose that all the desired sums are 1. In that case the limit matrix is doubly stochastic, and the structure of all doubly stochastic matrices is known (Birkhoff [1]). Thus it is possible to determine which cells have to be zero in the limit. Sinkhorn and Knopp [14] give necessary and sufficient conditions for convergence in the doubly stochastic case and show further that only those cells converge to zero that have to do so. That paper forms the starting point for the present one. In 1972 Sinkhorn [12] proved that in the same case one could set all cells that tended to zero equal to zero at the start without changing the limit. The corresponding theorem, proved differently, plays an important role in this paper.

After the paper by Sinkhorn and Knopp the attention of pure mathematicians shifted away from the ISP itself and towards the concept of diagonal equivalence. In 1968 Brualdi [2] gave necessary and sufficient conditions for the existence of a matrix with a given pattern of non-zero cells and given row and column sums. There are several papers, of greater or lesser generality, that use these conditions to show that if the initial matrix A has a suitable pattern, then there exists a unique matrix diagonally equivalent to A with the desired row and column sums (see [3], [5], [7], [8], [9] and [13]). These proofs are essentially topological in nature, using fixed point or optimum theorems. The norms they introduce are closely related to the auxiliary functions defined in Sinkhorn and Knopp and in this paper.

We prove here that a necessary and sufficient condition for the convergence of the ISP is that there exists a matrix with the given totals and zeros at least everywhere that the initial matrix has zeros. Such a matrix is used explicitly in the proof, rather than Brualdi's conditions. In the course of the proof we also show that the minimum possible number of cells tend to zero, and that if these cells are set to zero initially, then the process will converge to the same limit, which will then be diagonally equivalent to the new initial matrix. These results constitute Theorem 1. In a second section they are extended to the case when certain column sums are left unspecified (Theorem 2).

The author would like to express his gratitude to Datum e.V. of Bonn for suggesting and supporting the research contained in this paper, and to the referee for his helpful comments.

Received 22 January, 1979; revised 7 May, 1979.

All matrices in this paper are understood to be real and non-negative.

Two concepts are required continually in what follows, so we repeat their definitions and elementary properties here.

Two matrices, A and B, are said to be *diagonally equivalent* if there exist invertible diagonal matrices X and Y such that $B = XAY$. The diagonal elements of X are labelled x_i and called row multipliers, those of Y are labelled y_j and called column multipliers.

The *pattern* of a matrix A is a bipartite graph with a node for each row and each column and an edge connecting row node i to column node j whenever $a_{ij} \neq 0$. A matrix B is said, by abuse of language, to have a *partial pattern* of A if its pattern is a partial graph of the pattern of A (same nodes, subset of the edges). This is equivalent to $a_{ij} = 0 \Rightarrow b_{ij} = 0$. Diagonally equivalent matrices obviously have the same pattern. The matrix A is called *connected* if its pattern is a connected graph. Any matrix is the direct sum of its connected components. In a graph two nodes are called *neighbours* if there is an edge connecting them. By abuse of language we extend this terminology to rows and columns (and even row multipliers and column multipliers).

We now state Theorem 1 formally.

THEOREM 1. *Let a matrix A and desired row and column sums r_i, $c_j \neq 0$ be given. A necessary and sufficient condition for the* ISP *to converge to a matrix Q with these sums is the existence of a matrix B with partial pattern of A and the desired sums. Furthermore the pattern of Q is the maximal partial pattern of A for which such a matrix B exists. If A' is the matrix derived from A by setting cells outside that pattern to zero, then A' is diagonally equivalent to Q and the* ISP *starting with A' converges to Q.*

Our first proposition, the proof of which is a straightforward generalization of one in Sinkhorn [11] for the case of positive matrices, will be used to establish the uniqueness of the limit matrices in the ISP.

PROPOSITION 1. *If A and B are diagonally equivalent and have the same row and column sums, then $A = B$.*

Proof. We may assume that A and B are not 0 and, since they have the same pattern, we may prove the proposition for each of their components. So we assume that A is connected and that $B = XAY$; we designate the sums by r_i and c_j and may also assume that none of these is zero. Then

$$r_i = \sum_j b_{ij} = x_i \sum_j a_{ij} y_j \leqslant x_i r_i (\max y_j).$$

So $x_i \geqslant (\max y_j)^{-1}$ and equality occurs only if y_j is maximal for all neighbours of x_i. Similarly we obtain $y_j \leqslant (\min x_i)^{-1}$ with an analogous condition for equality. It follows that $\min x_i = (\max y_j)^{-1}$. Now let $x_1 = \min x_i$, say; then $y_j = \max y_j$ for all neighbours of x_1, hence $x_i = \min x_i$ for all neighbours of these y_j etc. From the connectedness of A it follows inductively that all $x_i = \min x_i$ and all $y_j = \max y_j$ and hence that $B = A$.

Before we start the proof of Theorem 1 itself we state a lemma on positive real numbers.

LEMMA 1. *Let $a_1, \ldots, a_t$ be positive real numbers and for $x_1, \ldots, x_t$ variable positive real numbers define*

$$f(x_1, \ldots, x_t) = \prod_i (a_i/x_i)^{a_i}.$$

Then for t-tuples (x_i) such that $\sum_i x_i = \sum_i a_i$, we have $f(x_1, \ldots, x_t) \geqslant 1$ and $f \to 1$ if and only if $x_1 \to a_i$ for each i.

This lemma can be proved using the convexity of the log function, or by Lagrange's method, or it can be derived from the generalized geometric-algebraic mean inequality.

The proof of Theorem 1 is broken into a series of propositions. We let the sequence of matrices generated by the ISP be $A^{(m)}$, where $A^{(2n)}$, $(n > 0)$, has column sums c_{jn} and the desired row sums, while $A^{(2n+1)}$ has row sums r_{in} and the desired column sums. Further $A^{(2n)} = X_n A Y_n$ and $A^{(2n+1)} = X_n A Y_{n+1}$. Since the sequence is bounded it must have limit matrices (of subsequences).

PROPOSITION 2. *If a matrix B as in Theorem 1 exists, then all limit matrices of the* ISP *have the desired row and column sums.*

Proof. We consider the functions

$$f_n = \prod_i x_{in}^{r_i} \prod_j y_{jn}^{c_j}$$

and

$$g_n = \prod_i x_{in}^{r_i} \prod_j y_{jn+1}^{c_j}.$$

By the construction of the multipliers we have

$$g_n/f_n = \prod_j (y_{jn+1}/y_{jn})^{c_j} = \prod_j (c_j/c_{jn})^{c_j}.$$

For $n > 0, \sum_j c_{jn} = \sum_j c_j$, so it follows from Lemma 1 that $g_n \geqslant f_n$ and that $g_n/f_n \to 1$ if and only if $c_{jn} \to c_j$ for all j. The analogous argument holds for f_{n+1}/g_n. Hence the conclusion will be reached if we can show that the sequence f_n is bounded above.

Now $x_{in} y_{jn} = a_{ij}^{(2n)}/a_{ij}$ for $a_{ij} \neq 0$, and so this product is bounded above, because the elements $a_{ij}^{(2n)}$ are bounded by r_i (if $n > 0$). A possible bound is thus $L = (\max r_i)/(\min a_{ij})$, where the minimum is taken over the non-zero elements of A.

At this point we exploit the existence of B. Note that the row and column sums of B are r_i and c_j respectively. Thus

$$f_n = \prod_{ij} (x_{in} y_{jn})^{b_{ij}}.$$

Furthermore, if $b_{ij} \neq 0$, then $a_{ij} \neq 0$, and hence $x_{in} y_{jn} \leqslant L$. So $f_n \leqslant L^d$, where $d = \sum_{ij} b_{ij}$ and f_n is bounded as desired.

PROPOSITION 3. *Under the same hypothesis, the pattern of any limit matrix is intermediate between those of B and A.*

Proof. With L as in Proposition 2, $(x_{in}y_{jn})^{b_{ij}}L^{(d-b_{ij})} \geqslant f_n \geqslant f_1$. So if $b_{ij} \neq 0$, then $x_{in}y_{jn}$ is bounded away from 0. A similar argument holds for $x_{in}y_{jn+1}$. Thus $a_{ij}^{(m)}$ is also bounded away from 0 for all m. This proves the proposition.

Note that since the choice of B is free within the constraints it follows that all limit matrices must have the same pattern, which must be the maximum possible. Let A' be derived from A by setting all elements outside this pattern equal to zero, and leaving the others unchanged.

PROPOSITION 4. *Under the same hypothesis, any limit matrix of the* ISP *applied to A is diagonally equivalent to A'.*

Proof. Let G be the limit of the sequence $X'_nAY'_n$ (where the ′ signifies that this is some subsequence of the ISP sequence). Then G is also the limit of the sequence $X'_nA'Y'_n$. So the proposition follows directly from the following lemma, which is a generalization of one proved by Sinkhorn and Knopp [14] for the special case that G is doubly stochastic.

LEMMA 2. *Let $G = \lim X'_kHY'_k$ where the matrices X'_k and Y'_k are positive diagonal matrices. If G has the same pattern as H, then they are diagonally equivalent.*

Proof. We assume without loss of generality that H is connected. By the hypothesis $\lim_k (x'_{ik}y'_{jk})$ exists and is non-zero whenever $h_{ij} \neq 0$. Let $x_{ik} = x'_{ik}/x'_{1k}$ and $y_{jk} = y'_{jk}x'_{1k}$ for all i and j. Then $x_{ik}y_{jk} = x'_{ik}y'_{jk}$; so $X_kHY_k = X'_kHY'_k$.

Now $x_{1k} = 1$ for all k so $\lim_k x_{1k} = 1$ exists. Suppose that it has been proved, for some neighbour x_{ik} of y_{jk}, that $\lim_k x_{ik} = x_i$ exists and is non-zero. Then it follows that $y_j = \lim_k y_{jk} = \lim_k (x'_{ik}y'_{jk})/\lim_k x_{ik}$ exists and is also non-zero. Of course the same argument allows us to proceed from a column multiplier to a neighbouring row multiplier. Since H is connected, it follows that all sequences (x_{ik}) and (y_{jk}) converge, and the desired conclusion follows.

It is clear that these propositions and the remark that $A'' = A'$ establish Theorem 1 completely. We now turn to the case in which some of the column totals are left unspecified. It is easy to see that we can add these columns together to form a single column and separate them at any desired stage of the process. We therefore assume that only one column total is unspecified, namely c_1. Now the sum of the column totals must equal the sum of the row totals in any matrix and from this fact we can calculate the only value of c_1 for which the problem is feasible. We can then insert this total and apply the standard ISP and Theorem 1. However, we would also like to establish what happens if we apply a modified ISP in which all column multipliers y_{1n} are set to one. We will prove the following theorem.

THEOREM 2. *A necessary and sufficient condition for the convergence of the modified* ISP (MISP) *is the existence of a matrix B of partial pattern of the initial matrix satisfying the marginal conditions. In that case the limit of* MISP *is the same as that of* ISP.

To establish Theorem 2 we follow the pattern of the proof of Theorem 1. It is clear that once we have established the equivalents of Propositions 2 and 3, Proposition 4 and the main part of the theorem will follow as before, since the marginal totals do not enter into that part of the proof. Proposition 1 then establishes the last part of the theorem.

Before starting the proof we note that the elements of the first column of $A^{(2n+1)}$ are now bounded by the row totals, while all other elements are bounded exactly as before. (We are using the same notation as in Theorem 1, but it is now understood to apply to the MISP.)

PROPOSITION 5. *Under the hypothesis of Theorem 2, any limit matrix of the modified procedure satisfies the marginal conditions.*

Proof. We define f_n and g_n as in Proposition 2. However, it is no longer necessarily true that $g_n \geqslant f_n$, since $g_n/f_n = \prod_{j \geqslant 2} (c_j/c_{jn})^{c_j}$, and $\sum_{j \geqslant 2} c_{jn} = \sum_{j \geqslant 2} c_j$ need not hold. Thus Lemma 2 cannot be applied, but it does give a lower bound for g_n/f_n:

$$g_n/f_n \geqslant \left(\sum_{j \geqslant 2} c_j \Big/ \sum_{j \geqslant 2} c_{jn} \right)^{\sum_{j \geqslant 2} c_j},$$

and this lower bound is approached only if the ratios c_j/c_{jn} $(j \geqslant 2)$ all approach the same value.

Again, since $\sum_i r_{in} = c_{1n} + \sum_{j \geqslant 2} c_j$, it is not necessarily true that $f_{n+1} \geqslant g_n$, and we have the lower bound

$$f_{n+1}/g_n \geqslant \left(\sum_i r_i \Big/ \sum_i r_{in} \right)^{\sum_i r_i}$$

with an analogous condition for the bound to be approached.

However it is true that $f_{n+1} \geqslant f_n$. For if we put $c_{1n} = a$, $c_1 = a+u$, $\sum_{j \geqslant 2} c_j = b$, then $\sum_i r_i = a+b+u$, $\sum_i r_{in} = a+b$ and $\sum_{j \geqslant 2} c_{jn} = b+u$. The statement then follows from the following lemma.

LEMMA 3. *If a and b are positive numbers and $u > -b$, then*

$$\left(\frac{a+b+u}{a+b} \right)^{(a+b+u)} \left(\frac{b}{b+u} \right)^{b} \geqslant 1,$$

and it approaches 1 *only as u approaches* 0.

The lemma can easily be proved by differentiation.

Now if $f_{n+1}/f_n \to 1$, it follows that $u \to 0$, so that the lower bounds for g_n/f_n and f_{n+1}/g_n approach 1. But then g_n/f_n and f_{n+1}/g_n must approach their lower bounds and so the ratios c_j/c_{jn} $(j \geqslant 2)$ and r_i/r_{in} must approach constants (in terms of i and j). But since $c_{1n} \to 1$, these constants must be 1.

To complete the proof of the proposition it therefore remains to establish an upper bound for f_n. But this can be done precisely as in Proposition 2.

Remark. This argument cannot be applied to g_{n+1}/g_n, as there is no similar direct link between the errors of $\sum_i r_{in}$ and $\sum_{j \geqslant 2} c_{jn+1}$.

PROPOSITION 6. *Under the hypothesis of Theorem 2 any limit matrix of the modified procedure has a pattern intermediate between that of B and that of A.*

Proof. For limits of the sequence $(A^{(2n)})$ this can be established just as in Proposition 3. For the odd matrices a little extra work is required. We note that we have established that $\lim_n(c_j/c_{jn}) = 1$ for all j. But remarking that $c_j/c_{jn} = a_{ij}^{(2n+1)}/a_{ij}^{(2n)}$ for all $j \geqslant 2$, it follows that if a sequence of elements $a_{ij}^{(2n+1)}$ converges to 0, the same must be true for the corresponding sequence $a_{ij}^{(2n)}$ and *vice versa.* Hence the patterns of limit matrices of $(A^{(2n+1)})$ are the same as those of limit matrices of $(A^{(2n)})$. This concludes the proof of the proposition.

Since the proofs of Propositions 1 and 4 remain valid (as we have already remarked), this establishes Theorem 2. We mention that numerical examples suggest that the convergence of MISP is significantly slower than that of the original ISP, and they also suggest that convergence still occurs when certain row and column sums are simultaneously left unspecified. However the error analysis in Proposition 5 is too crude to establish the behaviour of f_n or g_n in that case.

References

1. G. Birkhoff, "Tres observaciones sobre el algebra lineal", *Rev. Univ. Nacional de Tacuman* A, 4 (1946), 147–150.
2. R. A. Brualdi, "Convex sets of non-negative matrices", *Canad. J. Math.*, 20 (1968), 144–157.
3. R. A. Brualdi, S. V. Parter and H. Schneider, " The diagonal equivalence of a non-negative matrix to a stochastic matrix", *J. Math. Anal. Appl.*, 16 (1966), 31–50.
4. W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the marginal totals are known", *Ann. Math. Stat.*, 11 (1940), 427–444.
5. D. Z. Djokovic, "Note on non-negative matrices", *Proc. Amer. Math. Soc.*, 25 (1970), 80–82.
6. S. E. Fienberg, "An iterative procedure for estimation in contingency tables", *Ann. Math. Stat.*, 41 (1970), 907–917.
7. D. London, "On matrices with doubly stochastic pattern", *J. Math. Anal. Appl.*, 34 (1971), 648–652.
8. A. W. Marshall and I. Olkin, "Scaling matrices to achieve specified row and column sums", *Numer. Math.*, 12 (1968) 83–90.
9. M. V. Menon, "Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with prescribed row and column sums", *Canad. J. Math.*, 20 (1968), 225–232.
10. M. V. Menon and H. Schneider, "The spectrum of a non-linear operator associated with a matrix", *Linear Algebra Appl.*, 2 (1969), 321–334.
11. R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices", *Ann. Math. Stat.*, 35 (1964), 876–879.
12. R. Sinkhorn, "Continuous dependence on A in the D_1AD_2 theorems", *Proc. Amer. Math. Soc.*, 32 (1972), 395–398.
13. R. Sinkhorn, "Diagonal equivalence to matrices with prescribed row and column sums II", *Proc. Amer. Math. Soc.*, 45 (1974), 195–198.
14. R. Sinkhorn and P. Knopp, "Concerning non-negative matrices and doubly stochastic matrices", *Pacific J. Math.*, 21 (1966), 343–348.

Imperial College,
London, S.W.7.