Universität Augsburg

Institut für
Mathematik

Friedrich Pukelsheim, Bruno Simeone

On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L_1-Error Analysis

Preprint Nr. 05/2009 — 17. März 2009
Institut für Mathematik, Universitätsstraße, D-86135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:

Institut für Mathematik
Universität Augsburg
86135 Augsburg
`http://www.math.uni-augsburg.de/forschung/preprint/`

ViSdP:

Friedrich Pukelsheim
Institut für Mathematik
Universität Augsburg
86135 Augsburg

ON THE ITERATIVE PROPORTIONAL FITTING PROCEDURE: STRUCTURE OF ACCUMULATION POINTS AND L_1-ERROR ANALYSIS

Friedrich Pukelsheim and Bruno Simeone

Universität Augsburg and Sapienza Università di Roma

A new analysis of the Iterative Proportional Fitting procedure is presented. The input data consist of a nonnegative matrix, and of row and column marginals. The output sought is a biproportional fit, that is, a scaling of the input matrix by means of row and column divisors so that the scaled matrix has row and column sums equal to the input marginals. The IPF procedure is an algorithm alternating between the fitting of rows and columns. The structure of its accumulation points is explored in detail. The progress of the algorithm is evaluated through an L_1-error function measuring the deviation of current row and column sums from target marginals. A formula is obtained, of max-flow min-cut type, to calculate the minimum L_1-error directly from the input data. If the minimum L_1-error is zero, the IPF procedure converges to the unique biproportional fit. Otherwise, it eventually oscillates between various accumulation points.

1. Introduction. We present a novel, L_1-based analysis of the Iterative Proportional Fitting (IPF) procedure. The IPF procedure is an algorithm for scaling rows and columns of a given $k \times \ell$ *weight matrix* $A = ((a_{ij}))$ so that the resulting, row-wise and column-wise scaled output matrix $B = ((b_{ij}))$ achieves row sums equal to a prespecified vector of *row marginals*, $r = (r_1, \ldots, r_k)$, and column sums equal to a prespecified vector of *column marginals*, $s = (s_1, \ldots, s_\ell)$. We assume all weights to be nonnegative, $a_{ij} \geq 0$, and all marginals to be positive, $r_i > 0$ and $s_j > 0$. No row nor column of A is allowed to vanish.

The problem has a continuous variant, the *biproportional fitting problem*, and a discrete variant, the *biproportional apportionment problem*. In the continuous variant, the entries of the matrix B sought are permitted to be any nonnegative real numbers, $b_{ij} \geq 0$. The output B is called a *biproportional fit*, of the weight matrix A to the row marginals r and to the column marginals s. The IPF procedure iteratively calculates scaled matrices $A(t) = \big((a_{ij}(t))\big)$, where for odd steps all row sums are matching, $a_{i+}(t) = r_i$ for all $i \leq k$, while for even steps the column sums match, $a_{+j}(t) = s_j$ for all $j \leq \ell$. If a biproportional fit B exists, the sequence of scaled matrices $A(1), A(2), \ldots$ converges to B.

AMS 2000 subject classification. 62H17; 62P25.

Key words and phrases. Biproportional fitting, biproportional apportionment, matrix scaling, RAS method, entropy, alternating scaling algorithm.

The discrete variant of the problem restricts the entries of B to be nonnegative integers, $b_{ij} \in \{0, 1, 2, \ldots\}$. Row and column marginals must then be integers, of course. Now the output matrix B is called a *biproportional apportionment*, according to the given weight matrix A and the prespecified row marginals r and column marginals s. The analogous solution method for the discrete problem is the Alternating Scaling (AS) algorithm. At step t it produces a matrix $A(t)$ with entries $a_{ij}(t)$ being scaled as well as rounded, in order to comply with the pertinent integer restrictions. The AS algorithm aims to solve the discrete problem variant. Due to the occurrence of ties there are (rare) instances, however, where a biproportional apportionment B exists while the AS algorithm fails to converge to it and stalls; see Example 8.1 in Gaffke and Pukelsheim (2008).

Our research arose from the desire to better understand the interplay between the continuous IPF procedure, and the discrete AS algorithm. The present paper focuses on the continuous fitting problem. However, our major tool, the L_1-error function

$$f(t) = \frac{1}{2} \sum_{i \le k} \left| a_{i+}(t) - r_i \right| + \frac{1}{2} \sum_{j \le \ell} \left| a_{+j}(t) - s_j \right|,$$

is borrowed from Balinski and Demange's (1989) enquiry into the discrete apportionment problem. In the discrete case the error function $f(t)$ is quite suggestive, simply counting along rows and columns how many units are wrongly allocated at step t. For the continuous problem the L_1-error $f(t)$ is, at first glance, just one out of many ways to assess lack of fit. At second glance it is a most appropriate way, as this paper shows.

1.1. The literature on biproportional fitting. The continuous biproportional fitting problem is the senior member of the problem family. It has created an enormous body of literature of which we review only the papers that influenced the present research. The term *IPF procedure* prevails in Statistics, see Fienberg and Meyer (2006), or Speed (2005); we follow those leads. Some Statisticians prefer *matrix raking*, such as Fagan and Greenberg (1987). In Operations Research and Econometrics the label *RAS method* is popular, pointing to a (diagonal) matrix R of row multipliers, the weight matrix A, and a (diagonal) matrix S of column multipliers, as mentioned already by Bacharach (1965, 1970). Computer scientists prefer the term *matrix scaling*, as in Rote and Zachariasen (2007).

The IPF procedure was popularized by Deming and Stephan (1940), though there are earlier papers where the idea was used, see Fienberg and Meyer (2006). Deming and Stephan (page 440) recommend terminating iterations when *the table reproduces itself*, that is, in our terminology, when the scaled matrices $A(t)$ and $A(t+1)$ get close to each other.

This distance is what is measured by the L_1-error function $f(t)$, see our Lemma 1. Deming and Stephan successfully advocated the merits of the algorithm, but were somewhat led astray in its analysis, as communicated by Stephan (1942).

Brown (1959) proposed a proof of convergence which Ireland and Kullback (1968) found to lack rigor. The latter authors established convergence by relating the IPF procedure to the minimum entropy solution. Csiszár (1975, page 155) noted that their convergence proof was incomplete, and that the generalization to measure spaces by Kullback (1968) suffered from a similar flaw. Csiszár (1975) salvaged the entropy approach, and Rüschendorf (1995) established its extension to general measure spaces.

Despite of all the emphasis on entropy, the ultimate arguments of Ireland and Kullback (1968), equations (4·32) and (4·33) on page 185, substitute convergence of entropy by convergence in L_1, referring to a result of Kullback (1966). Also Bregman (1967) starts out with entropy, and then uses the L_1-error function. Here we dispose of the entropy detour, and use L_1 from start to finish. Ireland and Kullback (1968, page 184) prove that the entropy criterion decreases monotonically, as does the likelihood function of Bishop, Fienberg and Holland (1975, page 86), and our L_1-error function (Lemma 1).

Over time authors replaced entropy by related criteria, see the list on page 376 of Kalantari, Lari, Ricca and Simeone (2008). Marshall and Olkin (1968) and Macgill (1977) minimized a quadratic objective function. Pretzel (1980) used a weighted geometric mean. We find that Pretzel's is perhaps the most elegant, and certainly one of the shortest proofs of convergence of the IPF procedure. The only drawback is that Pretzel builds on a necessary and sufficient condition for the existence of a solution that, in our exposition, comes towards the end, as condition (2) in our Theorem 4.

The question when a biproportional fit exists generated a wealth of papers by itself, such as Brualdi, Parter and Schneider (1966), and Schneider (1990). Many of them are oriented towards network and graph theory; we make use of such arguments in Section 6. Moreover, the formula for the minimum L_1-error in Theorem 3 is so easy to evaluate, prior to starting the IPF procedure, because it is akin to treating the issue as a transportation problem. Rachev and Rüschendorf (1998) present an in-depth development of measure-theoretic mass transportation problems, and we tend to believe that there are more interrelations than we have been able to identify.

An entirely different route was opened by Fienberg (1970) who embedded the IPF procedure into the geometry of the manifold of constant interaction in a $(k\ell - 1)$-dimensional simplex of reference. Fienberg worked under the assumption that all input weights are positive, $a_{ij} > 0$. He pointed out (page 915) that the extension to problems involving zero weights is *quite complex*, which is attested to by much of the literature. Ireland and Kullback's (1968, page 182) plea to assume positive weights *to simplify the argument* is a friendly understatement, unless it is meant to be the utter truth.

Another approach, keeping as close to calculus as possible, was championed early on by Bacharach (1965, 1970), and by Sinkhorn (1964, 1966, 1967, 1972, 1974) and Sinkhorn and Knopp (1967). Much of the present paper is owed to Bacharach. Michael Owen Leslie Bacharach (b. 1936, d. 2002) was an Oxford econometrician of some renown. In 1965 he earned a PhD degree in Mathematics from Cambridge; his thesis was published as Bacharach (1965), and became Section 4 of Bacharach (1970). Richard Dennis Sinkhorn (b. 1934, d. 1995) received his Mathematics PhD in 1962 from the University of Wisconsin–Madison, with a thesis entitled *On Two Problems Concerning Doubly Stochastic Matrices*. Throughout his career he served as a Mathematics professor with the University of Houston. Though contemporaries, neither of the two ever quoted the other.

1.2. The literature on biproportional apportionment. The discrete biproportional apportionment problem is the junior member of the problem family, first put forward by Balinski and Demange (1989), see also Balinski and Rachev (1997). The operation of rounding scaled quantities to integers sounds most attractive for the statistical analysis of frequency tables, as remarked by Wainer (1998) and Pukelsheim (1998). It disposes of any disclaimers that the adjusted figures are *rounded off, hence when summed may occasionally disagree a unit or so*, as in Table I of Deming and Stephan (1940, page 433). When the task is to apportion 100 percentage points, as in Table 3.6-4 of Bishop, Fienberg and Holland (1975, page 99), the method would not stop short with 99 percent only. However, Balinski was motivated not by discrete multivariate statistics, but by the perceived use of such methods for parliamentary elections in proportional representation systems.

The task of allocating seats of a parliamentary body to political parties does not tolerate any disclaimers excusing residual rounding errors. Whatever apportionment method is used, it must meticulously account for every single seat. This is achieved by biproportional apportionment methods. In 2003, the Swiss Canton of Zurich amended their Electoral Law to adopt the *biproportional divisor method with standard rounding*, see Pukelsheim and Schuhmacher (2004), and Balinski and Pukelsheim (2006). The Canton of Schaffhausen and the Canton of Aargau followed suit in 2007. The method is attractive also for other countries, as worked out by Pennisi (2006) for Italy; Zachariassen and Zachariasen (2006) for the Farœ Islands; and Ramírez, Pukelsheim, Palomares and Martínez (2008) for Spain.

One of the authors (FP) had the privilege of advising the Zurich Parliament on the matter. He felt it inappropriate to persuade politicians that the new method is akin to minimizing entropy, or that it is justified through differential geometry of smooth manifolds in high-dimensional simplexes. The procedure simply does what proportional apportionment calls for: Scale and round! Scaling within electoral districts (rows) achieves proportionality among the parties running in that district. Scaling within parties (columns) secures district lists of any party to be handled proportionally. The final rounding step is felt inevitable, as deputies come in whole numbers and not in fractions. In the end the biproportional divisor method with standard rounding won overwhelming political support.

That it also won strong *administrative* support is a victory of the IPF procedure and its discrete sibling, the AS algorithm, enabling the officials to easily calculate district divisors and party divisors. Once suitable divisors are publicized all voters can double-check the outcome. They only need to take the vote count of the party of their choice in their district, divide it by the respective district and party divisors, and round the result to the nearest seat number. A computer program for carrying out the apportionment is provided at `www.uni-augsburg.de/bazi`, see Pukelsheim (2004), Joas (2005), Maier and Pukelsheim (2007). The user may choose to run the AS algorithm, the Tie-and-Transfer (TT) algorithm of Balinski and Demange (1989), or hybrid combinations of the two. The performance of the algorithms is studied by Maier, Zachariassen and Zachariasen (2009).

In the electoral application the weight matrix A consists of vote counts, and the issue of zero weights gains real import. All too often there exists a party j not campaigning in some district i, and hence entering into the final results with $a_{ij} = 0$. The analysis may no longer be simplified by assuming all weights to be positive. Zero weights must be dealt with, even if the labor they entail becomes quite complex.

1.3. Section overview. We give a brief overview of the sections that follow. Section 2 introduces limit biproportional fits, and direct biproportional fits. If it exists, a limit biproportional fit is unique (Theorem 1). If a limit fit is connected, it is direct (Theorem 2). Direct biproportional fits are called 'interior' solutions by Bacharach (1970, page 45), whereas limit fits that are not direct are termed 'boundary' solutions.

Section 3 describes the IPF procedure, with its sequence of scaled matrices $A(t)$. Even steps generate incremental row divisors $\rho_i(t)$, odd steps create incremental column divisors $\sigma_j(t)$. They are likelihood ratios, of current sums relative to target marginals. Lemma 1 shows that the ensuing L_1-error function $f(t)$ is nonincreasing. Three examples illustrate that the error may decrease exponentially, or just linearly.

Section 4 scrutinizes the accumulation points of the sequence of scaled matrices $A(t)$, $t \geq 1$. Lemma 2 establishes a chain of inequalities between minimum and maximum incremental row and column divisors, implicit in the works of Bacharach, Sinkhorn, and Pretzel. Minimum and maximum incremental divisors are found to move monotonically towards each other. Without loss of generality we consider limit matrices along subsequences of even steps, whence their columns are matching and all remaining errors originate from their rows. Lemma 3 shows that the structure of an accumulation point B is contingent on the decomposition into its connected components.

Section 5 investigates the L_1-error function $f(t)$. Being nonnegative, the error is trivially bounded by zero. Lemma 4 exhibits lower bounds that are potentially tighter, $r_I - s_{J_A(I)}$, where r_I is the partial marginal sum over an arbitrary row subset I, and $s_{J_A(I)}$ is the partial marginal sum over the set $J_A(I)$ of columns connected in A to I. Our main result, Theorem 3, states that one of these lower bounds is sharp,

$$\lim_{t\to\infty} f(t) = \max_{I\subseteq\{1,\ldots,k\}} \Big(r_I - s_{J_A(I)} \Big) = r_{U_B} - s_{J_A(U_B)},$$

where the set U_B designates the underweighted rows of an even-step accumulation point B. Most of the proof is occupied by the case $U_B \neq \emptyset$ when some rows of B do not match their marginals, that is, when the minimum L_1-error stays positive.

Section 6 exploits the formula for the minimum L_1-error to re-derive the well-known necessary and sufficient conditions for the convergence of the IPF procedure. Theorem 4 addresses the existence of a limit biproportional fit in general. Theorem 5 focuses on the special instances when the limit matrix is disconnected. Theorem 6 considers the case when the limit matrix is connected, and hence the fit is direct. The latter has an immediate application to the very special case when all weights are positive; then the IPF procedures always converges, and the biproportional fit is direct.

1.4. Notation. The following notation turns out to be convenient. For a vector $r = (r_1, \ldots, r_k)$, the operators $+$, min, max indicate, respectively, the sum of its entries, $r_+ = \sum_{i\leq k} r_i$, a minimum entry, $r_{\min} = \min\{r_1, \ldots, r_k\}$, and a maximum entry, $r_{\max} = \max\{r_1, \ldots, r_k\}$. The partial sum of the marginals r_i over a row subset I is denoted by $r_I = \sum_{i\in I} r_i$. For column subsets J the notation extends to the partial sum of column marginals, $s_J = \sum_{j\in J} s_j$, as well as to the sum of entries in the $I \times J$ block of a matrix A, such as $a_{I\times J} = \sum_{i\in I} \sum_{j\in J} a_{ij}$. Sums over the empty set are taken to be zero, $r_\emptyset = s_\emptyset = a_\emptyset = 0$. The complement of any set $I \subseteq \{1, \ldots, k\}$ is indicated by a prime, $I' = \{1, \ldots, k\} \setminus I$. The set $J_A(I)$ of columns connected in A to I is defined in Section 5.

2. Biproportional Fits. Let A be a $k \times \ell$ matrix with nonnegative entries, $a_{ij} \geq 0$, called *weights.* No row nor column of A is allowed to vanish. We take $k \geq 2$ and $\ell \geq 2$, so as to deal with genuine matrices.

Let $r = (r_1, \ldots, r_k)$ and $s = (s_1, \ldots, s_\ell)$ be vectors with positive entries, $r_i > 0$ and $s_j > 0$, called *row marginals* and *column marginals.* Let h be the larger of the two component sums, $h = \max\{r_+, s_+\}$. In much of what follows they turn out to be the same, in which case we simply have $r_+ = s_+ = h$.

A $k \times \ell$ matrix $B = (\!(b_{ij})\!)$ is called a *limit biproportional fit* (*of the weight matrix A, to the row marginals r and to the column marginals s*) when there exist sequences of positive *row divisors* $x_i(1), x_i(2), \ldots$ and of positive *column divisors* $y_j(1), y_j(2), \ldots$ so that the elements of B are of the form

$$b_{ij} = \lim_{t \to \infty} \frac{a_{ij}}{x_i(t) y_j(t)}$$

and fit the marginals, $b_{i+} = r_i$ and $b_{+j} = s_j$, for all rows $i \leq k$ and for all columns $j \leq \ell$. Of course, a necessary condition for a limit biproportional fit to exist is that the marginals have the same component sums.

THEOREM 1 (UNIQUENESS). *If a limit biproportional fit exists, then it is unique.*

PROOF. We assume that there are two distinct limit biproportional fits, $B \neq C$, with divisor sequences $x_i(t)$ and $y_j(t)$ for B, and $u_i(t)$ and $v_j(t)$ for C. Since marginal sums coincide, unequal entries in some cells must be evened out through other cells.

The starting point is a column j_1 with $b_{i_1 j_1} > c_{i_1 j_1}$, for some row i_1. Then $a_{i_1 j_1}$ is positive, entailing $x_{i_1}(t) y_{j_1}(t) < u_{i_1}(t) v_{j_1}(t)$ for eventually all t. Next we inspect row i_1 which, being fitted, features an entry $b_{i_1 j_2} < c_{i_1 j_2}$ in some column j_2. This forces $a_{i_1 j_2} > 0$, and $u_{i_1}(t) v_{j_2}(t) < x_{i_1}(t) y_{j_2}(t)$ for eventually all t. Now column j_2 needs to be corrected, followed by a correction in some row i_2. The construction terminates in a column j_q and a row i_q, say, with corresponding inequalities $x_{i_q}(t) y_{j_q}(t) < u_{i_q}(t) v_{j_q}(t)$ and $u_{i_q}(t) v_{j_1}(t) < x_{i_q}(t) y_{j_1}(t)$. It is convenient to introduce $j_{q+1} = j_1$. Picking every other inequality to form products we obtain a contradiction, for eventually all t,

$$\prod_{p \leq q} x_{i_p}(t) y_{j_p}(t) < \prod_{p \leq q} u_{i_p}(t) v_{j_p}(t) = \prod_{p \leq q} u_{i_p}(t) v_{j_{p+1}}(t) < \prod_{p \leq q} x_{i_p}(t) y_{j_{p+1}}(t) = \prod_{p \leq q} x_{i_p}(t) y_{j_p}(t).$$

Therefore, if a first limit biproportional fit exists, there is no room for another one. □

A $k \times \ell$ matrix $B = ((b_{ij}))$ is called a *direct biproportional fit* when there exist row divisors $u_i > 0$ and column divisors $v_j > 0$ so that the elements of B are of the form

$$b_{ij} = \frac{a_{ij}}{u_i v_j}$$

and fit the marginals, $b_{i+} = r_i$ and $b_{+j} = s_j$, for all rows $i \le k$ and for all columns $j \le \ell$. Obviously there are fewer direct biproportional fits than there are limit biproportional fits. The question of how to identify them is answered by the notion of connectedness. A matrix $C = ((c_{ij}))$ is called *connected* when it is not disconnected.

A matrix $D = ((d_{ij}))$ is called *disconnected* when it admits a *disconnectedness partition into components P and Q*,

$$D = \begin{matrix} & J & J' \\ I & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}$}} \\ I' & & \end{matrix},$$

where I and I' are complementary row subsets, J and J' are complementary column subsets, and at least one of the index sets $I \times J'$ or $I' \times J$ is nonempty, $\big(I \times J'\big) \cup \big(I' \times J\big) \neq \emptyset$. In such a situation the entries d_{ij} vanish whenever $(i,j) \in \big(I \times J'\big) \cup \big(I' \times J\big)$, while the entries of the $I \times J$ matrix P and of the $I' \times J'$ matrix Q are nonnegative.

Theorem 2 (Connectedness). *Suppose A is connected and a limit biproportional fit B exists. Then B is a direct biproportional fit if and only if B is connected.*

Proof. For the direct part, let B be a direct biproportional fit. Then the zero patterns of A and B are the same. Since A is connected, so is B.

For the converse part, we apply a scanning process to consecutively identify divisors u_i for rows, and v_j for columns. We use that positive entries in B require positive entries in A. In step 1 we scan the first row by equipping it with divisor unity, $u_1 = 1$. In step 2 we scan the columns j with $b_{1j} > 0$, by defining $v_j = a_{1j}/\big(u_1 b_{1j}\big)$. In step 3 we scan the unscanned rows i where $b_{ij} > 0$ for some scanned column j, and set $u_i = a_{ij}/\big(b_{ij} v_j\big)$. In step 4 we turn back to scanning those columns that are still unscanned. Continuing back and forth, the process enlarges the scanned sets for at most $k + \ell$ steps. With I the set of scanned rows and J the set of scanned columns, B acquires block structure,

$$B = \begin{matrix} & J & J' \\ I & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}$}} \\ I' & & \end{matrix}.$$

Connectedness of B forces the sets I' and J' to be empty. That is, the process has scanned all rows and all columns, $I = \{1, \ldots, k\}$ and $J = \{1, \ldots, \ell\}$, and the fit B is direct. □

3. The IPF Procedure. The IPF procedure alternatingly scales rows and columns, to fit the weight matrix A to the row marginals r and to the column marginals s, as follows.

- Step 1 rescales the weights to fit rows, $a_{ij}(1) = a_{ij}/\rho_i(0)$, where the row divisors build on the input weights, $\rho_i(0) = a_{i+}/r_i$.
- Step 2 rescales the weights to fit columns, $a_{ij}(2) = a_{ij}(1)/\sigma_j(1)$, with column divisors calculated from the weights in the first step, $\sigma_j(1) = a_{+j}(1)/s_j$.
- Odd steps $t+1$ rescale the weights to fit rows, $a_{ij}(t+1) = a_{ij}(t)/\rho_i(t)$, with row divisors calculated from the previous even step, $\rho_i(t) = a_{i+}(t)/r_i$.
- Even steps $t+2$ rescale the weights to fit columns, $a_{ij}(t+2) = a_{ij}(t+1)/\sigma_j(t+1)$, with column divisors calculated from the previous odd step, $\sigma_j(t+1) = a_{+j}(t+1)/s_j$.

Divisors such as $\rho_i(t) = a_{i+}(t)/r_i$ are likelihood ratios, with the distribution to be fitted in the numerator, $a_{i+}(t)$, and the target distribution in the denominator, r_i. Of course, we could also work with the inverse quantities, multipliers.

Since no row and no column of A is allowed to vanish, the divisors always stay positive. Each step of the IPF procedure generates *incremental divisors*, either row divisors $\rho_i(0), \rho_i(2), \rho_i(4), \ldots$, or else column divisors $\sigma_j(1), \sigma_j(3), \sigma_j(5), \ldots$. Successive incremental divisors give rise to *cumulative divisors* $x_i(t)$ and $y_j(t)$, defined through

$$\rho_i(0)\,\rho_i(2)\,\rho_i(4)\,\cdots\,\rho_i(t-2) = x_i(t-1) = x_i(t),$$
$$\sigma_j(1)\,\sigma_j(3)\,\sigma_j(5)\,\cdots\,\sigma_j(t-1) = y_j(t) = y_j(t+1),$$

for $t = 2, 4, \ldots$. Adjoining $y_j(1) = 1$, the cumulative divisors are defined for all steps $t \geq 1$. In terms of the cumulative divisors, the scaled weights take the form

$$a_{ij}(t) = \frac{a_{ij}}{x_i(t)y_j(t)}.$$

The IPF procedure is said to converge to B when the scaled weight matrices $A(t) = \big(\!\big(a_{ij}(t)\big)\!\big)$, $t \geq 1$, have B as their limit, $\lim_{t\to\infty} A(t) = B$. The matrix B thus obtained visibly is the limit biproportional fit, of A to r and to s.

In case of convergence, the marginals r and s must have the identical component sums, namely the sum of all elements of B. By not assuming $r_+ = s_+$ in the present section and the next, we acquire a more comprehensive view of what may happen in the case of non-convergence.

We measure the step-wise progress of the IPF procedure by how much current row sums deviate from the prespecified row marginals, and by how much column sums deviate from the column marginals. To this end we introduce the L_1-*error function*

$$f(t) = \frac{1}{2} \sum_{i \le k} \Big| a_{i+}(t) - r_i \Big| + \frac{1}{2} \sum_{j \le \ell} \Big| a_{+j}(t) - s_j \Big|.$$

For odd steps t, rows match their prespecified marginals and the row error sum vanishes. In particular, $f(1) \le \frac{1}{2} \sum_{j \le \ell} \big(a_{+j}(1) + s_j \big) = \frac{1}{2}(r_+ + s_+) \le h$. For even steps t the column error sum is zero, as it is then the columns that attain their marginals. The factor $1/2$ accounts for any error appearing twice, as an overweight and as an underweight, as soon as we start assuming equal totals for the marginals, $r_+ = s_+$.

In terms of the L_1-error function $f(t)$, the scaled matrices $A(t)$ that are produced by the IPF procedure exhibit an increasingly better fit to the marginals r and s.

LEMMA 1 (MONOTONICITY). *The L_1-error function is nonincreasing, $f(t) \ge f(t+1)$ for every step $t \ge 1$.*

PROOF. For even steps t, all columns are fitted, $a_{+j}(t) = s_j$. The next step is odd and fits rows, $a_{i+}(t+1) = r_i$. For the transition to step $t+1$, consider some row i and its upcoming divisor, $\rho_i(t)$. If $\rho_i(t) \le 1$ then we have $a_{ij}(t) \le a_{ij}(t)/\rho_i(t) = a_{ij}(t+1)$, for all $j \le \ell$. If $\rho_i(t) \ge 1$ then $a_{ij}(t) \ge a_{ij}(t+1)$. Either way every row $i \le k$ is such that, for all $j \le \ell$, the nonzero differences $a_{ij}(t) - a_{ij}(t+1)$ are of the same sign. This yields

$$\begin{aligned} f(t) &= \frac{1}{2} \sum_{i \le k} \left| \sum_{j \le \ell} \Big(a_{ij}(t) - a_{ij}(t+1) \Big) \right| \\ &= \frac{1}{2} \sum_{i \le k} \sum_{j \le \ell} \Big| a_{ij}(t) - a_{ij}(t+1) \Big| = d\Big(A(t), A(t+1) \Big), \end{aligned}$$

say. The triangle inequality, applied within each column $j \le \ell$, establishes monotonicity,

$$\begin{aligned} d\Big(A(t+1), A(t) \Big) &= \frac{1}{2} \sum_{j \le \ell} \sum_{i \le k} \Big| a_{ij}(t+1) - a_{ij}(t) \Big| \\ &\ge \frac{1}{2} \sum_{j \le \ell} \left| \sum_{i \le k} \Big(a_{ij}(t+1) - a_{ij}(t) \Big) \right| = f(t+1). \end{aligned}$$

For odd steps t, monotonicity follows analogously. □

We find it instructive to illustrate the IPF procedure by example. Example 1 has the L_1-error function tending to zero exponentially fast, in Example 2 the speed is linear. In both examples the IPF procedure converges. In Example 3 the L_1-error function has limit two, and the IPF procedure fails to converge.

Example 1. Input and output are succinctly displayed through

$$\begin{array}{cccc} & 4 \quad 2 & \\ A = \begin{matrix} 2 \\ 4 \end{matrix} & \begin{pmatrix} 30 & 0 \\ 10 & 20 \end{pmatrix} & \begin{matrix} 1 \\ \frac{1}{3} \end{matrix} \\ & 15 \quad 30 & \end{array} \qquad \mapsto \qquad B = \begin{pmatrix} 2 & 0 \\ 2 & 2 \end{pmatrix}.$$

The 2×2 weight matrix A is bordered to the left by the prespecified row marginals, at the top, by the column marginals. The resulting row divisors are bordering A at the right, the column divisors, at the bottom. They yield the biproportional fit B shown.

The initial row divisors are $\rho_1(0) = 15$ and $\rho_2(0) = 15/2$, and make row sums match. Thereafter, for $t = 2, 4, \ldots$, we get

$$A(t-1) = \begin{pmatrix} 2 & 0 \\ 2\frac{2^{t-1}}{2^{t-1}+1} & 2\frac{2^{t-1}+2}{2^{t-1}+1} \end{pmatrix}, \qquad \begin{aligned} \sigma_1(t-1) &= \tfrac{2^t+1}{2^t+2}, \\ \sigma_2(t-1) &= \tfrac{2^{t-1}+2}{2^{t-1}+1}; \end{aligned}$$

$$A(t) = \begin{pmatrix} 2\frac{2^t+2}{2^t+1} & 0 \\ 2\frac{2^t}{2^t+1} & 2 \end{pmatrix}, \qquad \begin{aligned} \rho_1(t) &= \tfrac{2^t+2}{2^t+1}, \\ \rho_2(t) &= \tfrac{2^{t+1}+1}{2^{t+1}+2}. \end{aligned}$$

The L_1-error function takes values $f(t) = 2/(2^t+1)$, converging to zero exponentially fast. The IPF procedure evidently converges to the limit B displayed at the beginning.

Retrospectively we may calculate other, nicer divisors. With the first row divisor standardized to be unity, $u_1 = 1$, we obtain, one after another,

$$\frac{30}{u_1 v_1} = 2 \;\Rightarrow\; v_1 = 15, \qquad \frac{10}{u_2 v_1} = 2 \;\Rightarrow\; u_2 = \frac{1}{3}, \qquad \frac{20}{u_2 v_2} = 2 \;\Rightarrow\; v_2 = 30.$$

Visibly, then, B is seen to be connected, and to be a direct biproportional fit.

Example 2. The input keeps the weights A, but rearranges the marginals:

$$\begin{array}{cccc} & 3 \quad 3 & \\ A = \begin{matrix} 3 \\ 3 \end{matrix} & \begin{pmatrix} 30 & 0 \\ 10 & 20 \end{pmatrix} & \begin{matrix} 1 \\ t \end{matrix} \\ & 10 \quad \frac{20}{3}t^{-1} & \end{array} \qquad \mapsto \qquad B = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}.$$

The initial row divisors $\rho_1(0) = 10 = \rho_2(0)$ secure matching row sums. Thereafter, for $t = 2, 4, \ldots$, we get

$$A(t-1) = \begin{pmatrix} 3 & 0 \\ \frac{3}{t+1} & 3\frac{t}{t+1} \end{pmatrix}, \qquad \begin{aligned} \sigma_1(t-1) &= \tfrac{t+2}{t+1}, \\ \sigma_2(t-1) &= \tfrac{t}{t+1}; \end{aligned}$$

$$A(t) = \begin{pmatrix} 3\frac{t+1}{t+2} & 0 \\ \frac{3}{t+2} & 3 \end{pmatrix}, \qquad \begin{aligned} \rho_1(t) &= \tfrac{t+1}{t+2}, \\ \rho_2(t) &= \tfrac{t+3}{t+2}. \end{aligned}$$

The L_1-error tends to zero linearly, $f(t) = 3/(t+2)$. The IPF procedure converges to B.

The retrospective standardization of the cumulative divisors now fails to shortcut the limit character of B. Again starting with row divisor $u_1 = 1$, the identity $30/(u_1 v_1) = 3$ yields $v_1 = 10$. But then we get

$$\frac{10}{u_2(t)v_1} = 0 \;\Rightarrow\; u_2(t) = t, \quad \text{say,} \qquad \frac{20}{u_2(t)v_2(t)} = 3 \;\Rightarrow\; v_2(t) = \frac{20}{3}t^{-1}.$$

EXAMPLE 3. Our third example uses yet other marginals:

$$A = \begin{matrix} & \begin{matrix} 2 & 4 \end{matrix} \\ \begin{matrix} 4 \\ 2 \end{matrix} & \begin{pmatrix} 30 & 0 \\ 10 & 20 \end{pmatrix} \end{matrix}, \qquad \begin{aligned} \rho_1(0) &= \tfrac{15}{2}, \\ \rho_2(0) &= 15. \end{aligned}$$

Again the initial row divisors adjust row marginals. Thereafter, for $t = 2, 4, \ldots$, we get

$$A(t-1) = \begin{pmatrix} 4 & 0 \\ \frac{2}{2^t-1} & 2\frac{2^t-2}{2^t-1} \end{pmatrix}, \qquad \begin{aligned} \sigma_1(t-1) &= \tfrac{2^{t+1}-1}{2^t-1}, \\ \sigma_2(t-1) &= \tfrac{2^{t-1}-1}{2^t-1}; \end{aligned}$$

$$A(t) = \begin{pmatrix} 2\frac{2^{t+1}-2}{2^{t+1}-1} & 0 \\ \frac{2}{2^{t+1}-1} & 4 \end{pmatrix}, \qquad \begin{aligned} \rho_1(t) &= \tfrac{2^t-1}{2^{t+1}-1}, \\ \rho_2(t) &= \tfrac{2^{t+2}-1}{2^{t+1}-1}. \end{aligned}$$

The L_1-error is $f(t) = 2\frac{1}{1-2^{-(t+1)}}$, and converges to the limit two exponentially fast. The IPF procedure does not converge, but oscillates between two distinct accumulation points,

$$\lim_{t=1,3,\ldots} A(t) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \lim_{t=2,4,\ldots} A(t) = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}.$$

4. Accumulation points. Specifically, if for an even step t the IPF procedure comes out with all row divisors equal to unity, the matrix $A(t)$ is the limit sought. Columns match their marginals because the step t is even, and rows need no further adjustment because of unity divisors. The IPF procedure keeps reproducing $A(t)$, the limit B is reached.

Generally, the IPF procedure moves the incremental divisors towards each other, the smallest getting larger and the largest getting smaller, as detailed by the following interlacing inequalities. The smallest and the largest of the incremental divisors are denoted by $\rho_{\min}(t)$ and $\rho_{\max}(t)$, and by $\sigma_{\min}(t+1)$ and $\sigma_{\max}(t+1)$. We proceed in steps of two, starting with $t_0 = 2$ and hence with fitted columns. When A itself already comes with all columns fitted, which we indicate by writing $A(0)$ in place of A, we start from $t_0 = 0$.

LEMMA 2 (INEQUALITIES). (i) *For all even steps $t \geq t_0$ we have*

$$\rho_{\min}(t) \overset{(1)}{\leq} \frac{1}{\sigma_{\max}(t+1)} \overset{(2)}{\leq} \rho_{\min}(t+2) \leq \rho_{\max}(t+2) \overset{(3)}{\leq} \frac{1}{\sigma_{\min}(t+1)} \overset{(4)}{\leq} \rho_{\max}(t).$$

The four sequences of minimum and maximum incremental divisors of rows and of columns are convergent, and fulfill

$$\lim_{t=0,2,\dots} \rho_{\min}(t) = \frac{1}{\lim_{t=1,3,\dots} \sigma_{\max}(t)}, \qquad \lim_{t=0,2,\dots} \rho_{\max}(t) = \frac{1}{\lim_{t=1,3,\dots} \sigma_{\min}(t)}.$$

(ii) *If $\rho_{\min}(t) < \rho_{\max}(t)$ and $\rho_{\min}(t) = \rho_{\min}(t+2k-2)$, for some even step $t \geq t_0$, then A is disconnected.*

PROOF. (i) Let $t \geq t_0$ be even. For all rows $i \leq k$ and all columns $j \leq \ell$ we then have $a_{+j}(t) = a_{+j}(t+2) = s_j$ and $a_{i+}(t+1) = a_{i+}(t+3) = r_i$. This yields

$$1 = \frac{a_{+j}(t+2)}{s_j} = \frac{1}{s_j} \sum_{p \leq k} \frac{a_{pj}(t)}{\rho_p(t)\sigma_j(t+1)} \begin{cases} \leq \dfrac{1}{\rho_{\min}(t)\sigma_j(t+1)}, & (1j) \\ \geq \dfrac{1}{\rho_{\max}(t)\sigma_j(t+1)}; & (4j) \end{cases}$$

$$1 = \frac{a_{i+}(t+3)}{r_i} = \frac{1}{r_i} \sum_{q \leq \ell} \frac{a_{iq}(t+1)}{\rho_i(t+2)\sigma_q(t+1)} \begin{cases} \leq \dfrac{1}{\rho_i(t+2)\sigma_{\min}(t+1)}, & (3i) \\ \geq \dfrac{1}{\rho_i(t+2)\sigma_{\max}(t+1)}. & (2i) \end{cases}$$

Inequalities (1)–(4) in part (i) follow by forming maxima or minima over $j \leq \ell$ or $i \leq k$,

$$\rho_{\min}(t)\sigma_{\max}(t+1) \overset{(1)}{\leq} 1 \overset{(4)}{\leq} \rho_{\max}(t)\sigma_{\min}(t+1),$$
$$\rho_{\max}(t+2)\sigma_{\min}(t+1) \overset{(3)}{\leq} 1 \overset{(2)}{\leq} \rho_{\min}(t+2)\sigma_{\max}(t+1).$$

The inequalities justify the statements on the four limits. All of them come to lie in the interval $[\rho_{\min}(t_0), \rho_{\max}(t_0)]$.

(ii) Disconnected components are constructed by means of the row subsets $I(\tau)$ where the divisor is minimum, and the column subsets $J(\tau+1)$ where the divisor is maximum,

$$I(\tau) = \left\{ i \le k \,\middle|\, \rho_i(\tau) = \rho_{\min}(\tau) \right\},$$
$$J(\tau+1) = \left\{ j \le \ell \,\middle|\, \sigma_j(\tau+1) = \sigma_{\max}(\tau+1) \right\},$$

whenever τ is even. The first assumption, $\rho_{\min}(t) < \rho_{\max}(t)$, secures that the row subset $I(t)$ is proper, $I(t) \subsetneq \{1,\ldots,k\}$.

The second assumption expands into an equality string, $\rho_{\min}(t) = 1/\sigma_{\max}(t+1) = \rho_{\min}(t+2) = \cdots = \rho_{\min}(t+2k-4) = 1/\sigma_{\max}(t+2k-3) = \rho_{\min}(t+2k-2)$. We work our way in sets of three,

$$\rho_{\min}(t+z) = \frac{1}{\sigma_{\max}(t+z+1)} = \rho_{\min}(t+z+2),$$

with $z = 0, 2, \ldots, 2k-4$. For $i \in I(t+2)$ equality holds in $(2i)$, whence all $q \notin J(t+1)$ come with $a_{iq}(t+1) = 0$ and hence $a_{iq} = 0$. For $j \in J(t+1)$ equality obtains in $(1j)$, whence all $p \notin I(t)$ fulfill $a_{pj}(t) = 0$ and hence $a_{pj} = 0$. Any row $i \in I(t+2) \setminus I(t)$ would vanish, having $a_{ij} = 0$ for $j \notin J(t+1)$ as well as for $j \in J(t+1)$. But vanishing rows in A are not allowed, and so $I(t+2) \subseteq I(t)$.

The argument carries forward to build the chain

$$\{1,\ldots,k\} \neq I(t) \supseteq I(t+2) \supseteq \cdots \supseteq I(t+2k-4) \supseteq I(t+2k-2) \neq \emptyset.$$

As at most $k-2$ inclusions can be strict, somewhere on the way from $z = 0$ to $z = 2k-4$ the equality $I(t+z) = I(t+z+2)$ must emerge. Thus A is disconnected, of the form

$$A = \begin{array}{c} \\ I(t+z+2) \\ I(t+z)' \end{array} \begin{array}{c} \begin{array}{cc} J(t+z+1) & J(t+z+1)' \end{array} \\ \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \end{array}. \qquad \square$$

The scaled matrices $A(t)$ have nonnegative entries bounded by $h = \max\{r_+, s_+\}$, and hence stay in the compact set $[0,h]^{k\times\ell}$ for all steps t. This guarantees the existence of accumulation points. The issue is what they look like.

Lemma 3 (Accumulation Points). *Suppose the limit $B = \lim_{n\to\infty} A(t_n)$ along a subsequence of even steps t_n has connected components $B^{(1)}, \ldots, B^{(K)}$:*

$$B = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_K \end{array} \begin{array}{c} \begin{array}{cccc} J_1 & J_2 & \cdots & J_K \end{array} \\ \begin{pmatrix} B^{(1)} & 0 & \cdots & 0 \\ 0 & B^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B^{(K)} \end{pmatrix} \end{array}.$$

Then B has row sums $b_{i+} = \dfrac{s_{J_m}}{r_{I_m}} r_i$, for $i \in I_m$ and $m = 1, \ldots, K$.

When the IPF procedure is employed to fit B to the row marginals r and to the column marginals s, the divisors generated are $\alpha_i(0) = \frac{s_{J_m}}{r_{I_m}}$ for the rows and $\beta_j(1) = \frac{r_{I_m}}{s_{J_m}}$ for the columns, for all $i \in I_m$ and $j \in J_m$ and $m = 1, \ldots, K$. The resulting sequence of scaled matrices oscillates, $B(0) = B(z)$ and $B(1) = B(z+1)$, for all even z.

Proof. I. The even steps t_n pass on matching columns to B. No column vanishes, since $b_{+j} = s_j$. Nor vanishes any row, $b_{i+} = \lim_{n\to\infty} a_{i+}(t_n) = r_i \lim_{n\to\infty} \rho_i(t_n) \geq r_i \rho_{\min}(0) > 0$. Therefore we may apply the IPF procedure to the input matrix $B = B(0)$, to fit it to the row marginals r and and to the column marginals s.

For steps $z = 0, 2, \ldots$, the incremental divisors of $B(z)$ are denoted by $\alpha_i(z)$ and $\beta_j(z+1)$. They are linked to the incremental divisors of $A(t_n)$ through the limit relations

$$\lim_{n\to\infty} \rho_i(t_n + z) = \alpha_i(z), \qquad \lim_{n\to\infty} \sigma_j(t_n + z + 1) = \beta_j(z+1).$$

This follows from $\lim_{n\to\infty} \rho_i(t_n) = \lim_{n\to\infty} a_{i+}(t_n)/r_i = b_{i+}(0)/r_i = \alpha_i(0)$. Similarly we get $\lim_{n\to\infty} \sigma_j(t_n + 1) = \beta_j(1)$. Induction on $z = 0, 2, \ldots$ completes the argument.

II. We choose a sub-subsequence $z_p = t_{n_p}$, $p \geq 1$, of the subsequence t_n, $n \geq 1$, along which the scaled matrices $B(z_p)$ converge to a limit matrix C, say. Denoting the cumulative divisors of $B(z_p)$ by $u_i(z_p)$ and $v_j(z_p)$, we have

$$b_{ij} = \lim_{n\to\infty} \frac{a_{ij}}{x_i(t_n) y_j(t_n)}, \qquad c_{ij} = \lim_{p\to\infty} \frac{b_{ij}}{u_i(z_p) v_j(z_p)}.$$

Now we fix a row i and a column j for which c_{ij} is positive; then b_{ij} is positive, too, as is a_{ij}. Therefore the cumulative divisors converge to positive and finite limits,

$$\lim_{n\to\infty} x_i(t_n) y_j(t_n) = \frac{a_{ij}}{b_{ij}} \in (0, \infty), \qquad \lim_{p\to\infty} u_i(z_p) v_j(z_p) = \frac{b_{ij}}{c_{ij}} \in (0, \infty).$$

To investigate the interplay of the limits, we introduce the partial products

$$\Pi(t_n, z_p) = \prod_{z=0,2,\ldots,z_p-2} \rho_i(t_n+z)\sigma_j(t_n+z+1),$$

extending from $\rho_i(t_n)\sigma_j(t_n+1)$ through to $\rho_i(t_n+z_p-2)\sigma_j(t_n+z_p-1)$. We designate the limits with one argument tending to infinity and the other fixed by $\Pi(t_n,\infty) = \lim_{p\to\infty}\Pi(t_n,z_p)$, and by $\Pi(\infty,z_p) = \lim_{n\to\infty}\Pi(t_n,z_p)$.

For fixed n, we get $\Pi(t_n,\infty) = \lim_{N\to\infty} x_i(t_N)y_j(t_N)/\Pi(0,t_n) = (a_{ij}/b_{ij})/\Pi(0,t_n) \in (0,\infty)$. More than that, convergence transpires to be uniform in n,

$$\lim_{p\to\infty}\left(\sup_{n\ge1}\left|\Pi(t_n,z_p) - \frac{a_{ij}/b_{ij}}{\Pi(0,t_n)}\right|\right) = \left(\sup_{n\ge1}\frac{1}{\Pi(0,t_n)}\right)\lim_{p\to\infty}\left|\Pi(0,z_p) - \frac{a_{ij}}{b_{ij}}\right| = 0.$$

The supremum is finite because the subsequence $\Pi(0,t_n) = x_i(t_n)y_j(t_n)$ is convergent to $a_{ij}/b_{ij} \in (0,\infty)$, and the limit is zero because $z_p = t_{n_p}$ is a sub-subsequence of t_n.

For fixed p, we have $\Pi(\infty,z_p) = \prod_{z=0,2,\ldots,z_p-2}\alpha_i(z)\beta_j(z+1) = u_i(z_p)v_j(z_p)$. Because of uniform convergence the two transitions to the limit may be interchanged,

$$1 = \frac{a_{ij}/b_{ij}}{\lim_{n\to\infty}\Pi(0,t_n)} = \lim_{n\to\infty}\Pi(t_n,\infty) = \lim_{p\to\infty}\Pi(\infty,z_p) = \lim_{p\to\infty}u_i(z_p)v_j(z_p) = \frac{b_{ij}}{c_{ij}}.$$

This establishes $c_{ij} = b_{ij}$ whenever $c_{ij} > 0$. Since the column sums coincide, $c_{+j} = b_{+j} = s_j$ for all $j \le \ell$, so do the matrices as a whole, $C = B$. In other words, in the limit the scaled matrices $B(z_p)$ return to their starting point $B(0)$.

III. The grand problem, of fitting $B(0)$ to r and s, decomposes into the K smaller problems, of fitting $B^{(m)}(0)$ to $(r_i)_{i\in I_m}$ and $(s_j)_{j\in J_m}$, with $m = 1,\ldots,K$ fixed. We consider the minimum row divisors for the m-th subproblem, $\alpha^{(m)}_{\min}(z) = \min\{\alpha_i(z) \,|\, i \in I_m\}$ for even steps z. They are nondecreasing in z, by Lemma 2(i). We have just seen, however, that the subsequence $B(z_p)$ returns to the starting matrix $B(0)$. Therefore $\alpha^{(m)}_{\min}(z_p)$, too, returns to its starting value, $\alpha^{(m)}_{\min}(0)$. In view of monotonicity this forces the sequence to be constant, $\alpha^{(m)}_{\min}(0) = \alpha^{(m)}_{\min}(z)$ for all even z.

Connectedness of $B^{(m)}(0)$ now implies $\alpha^{(m)}_{\min}(0) = \alpha^{(m)}_{\max}(0)$, by Lemma 2(ii). Hence there is a constant μ with $\alpha_i(0) = \mu$ for all rows $i \in I_m$, giving row totals $b_{i+}(0) = \mu r_i$. Summing over all entries in $B^{(m)}(0)$ we end up with $\mu r_{I_m} = s_{J_m}$, and $\mu = s_{J_m}/r_{I_m}$.

With the structure of $B(0)$ being what it is, the sequence $B(z)$ of scaled matrices is seen to actually oscillate between $B(0)$ and $B(1)$. It is instantly verified that the IPF procedure behaves as is claimed in the assertion. □

The proof goes a long way to show that the accumulation point B has a rather simple structure. There is a shorter argument for the special case when the subsequence t_{n_p} is arithmetic, $t_{n_{p+1}} = t_{n_p} + q$, say. Along this subsequence we obtain

$$1 = \lim_{p\to\infty} \frac{x_i(t_{n_{p+1}})y_j(t_{n_{p+1}})}{x_i(t_{n_p})y_j(t_{n_p})} = \lim_{p\to\infty} \Pi(t_{n_p}, q) = u_i(q)v_j(q).$$

This yields $b_{ij}(0) = b_{ij}(q)$, and $B(0) = B(q)$. Thereafter the scaling process starts afresh, and repeats itself until $B(q) = B(2q)$. In the arithmetic case with increment q, the scaled matrices $B(z)$ return to their starting value $B(0)$ after period q, and not only in the limit.

5. Investigation of the L_1-Error. In the rest of the paper we assume equal component sums of the marginals, $r_+ = s_+ = h$. Then, if some incremental divisors are larger than unity, others must be smaller. Otherwise $\rho_p(t) > 1$ and $\rho_q(t) \geq 1$ for all $q \neq p$ entail a contradiction, $h = \sum_{i\leq k} a_{i+}(t+1) = \sum_{i\leq k}\sum_{j\leq \ell} a_{ij}(t)/\rho_i(t) < \sum_{j\leq \ell} a_{+j}(t) = h$.

The pattern of the positive entries in A now becomes essential. To monitor their occurrence we associate with every row subset I the *set of columns connected in A to I*, that is, the set of all columns containing a positive entry in some row of I,

$$J_A(I) = \{\, j \leq \ell \;\big|\; \exists i \in I\colon\ a_{ij} > 0 \,\}.$$

The extreme settings provide simple examples. If $I = \emptyset$ then we obtain $J_A(I) = \emptyset$. If $I = \{1,\ldots,k\}$ then we get $J_A(I) = \{1,\ldots,\ell\}$, since no row nor column of A vanishes. The complement $J_A(I)'$ embraces the columns with entries $a_{ij} = 0$ for all $i \in I$, whence we always have $a_{I\times J_A(I)'} = 0$.

We focus on steps t that are even, when current column sums attain the prespecified column marginals. For even steps t we introduce the set $U(t)$ of rows that are strictly underweighted, and the set $O(t)$ of rows that are strictly overweighted,

$$U(t) = \left\{\, i \leq k \;\middle|\; a_{i+}(t) < r_i \,\right\}, \qquad O(t) = \left\{\, i \leq k \;\middle|\; a_{i+}(t) > r_i \,\right\}.$$

LEMMA 4 (BOUNDS). *For all even steps t, every row subset $I \subseteq \{1, \dots, k\}$ with its set $J_A(I)$ of columns connected in A to I bound the L_1-error function from below via*

$$f(t) \ge r_I - s_{J_A(I)}.$$

Equality holds if and only if (i) *the set I contains all currently underweighted rows and no currently overweighted rows, $U(t) \subseteq I \subseteq O(t)'$, and* (ii) *the weight matrix A admits a disconnectedness partition into*

$$A = \begin{array}{c} \\ I \\ I' \end{array} \begin{array}{c} \begin{array}{cc} J_A(I) & J_A(I)' \end{array} \\ \left(\begin{array}{cc} P & 0 \\ 0 & Q \end{array} \right) \end{array}.$$

PROOF. Any two complementary row subsets I and I' satisfy

$$\sum_{i \in I} \Big(a_{i+}(t) - r_i \Big) + \sum_{i \in I'} \Big(a_{i+}(t) - r_i \Big) = \sum_{i \le k} \Big(a_{i+}(t) - r_i \Big) = a_{++}(t) - h = 0.$$

This gives rise to the identity $\sum_{i \in I'} \big(a_{i+}(t) - r_i \big) = \sum_{i \in I} \big(r_i - a_{i+}(t) \big)$.

Let step t be even, whence columns are fitted. The error function $f(t)$ satisfies

$$\begin{aligned} f(t) &= \frac{1}{2} \left(\sum_{i \in I} \Big| a_{i+}(t) - r_i \Big| + \sum_{i \in I'} \Big| a_{i+}(t) - r_i \Big| \right) \\ &\ge \frac{1}{2} \left(\sum_{i \in I} \Big(r_i - a_{i+}(t) \Big) + \sum_{i \in I'} \Big(a_{i+}(t) - r_i \Big) \right) \\ &= \sum_{i \in I} \Big(r_i - a_{i+}(t) \Big). \end{aligned}$$

Equality holds if and only if condition (i) applies.

We decompose the sum of weights,

$$\begin{aligned} \sum_{i \in I} \Big(r_i - a_{i+}(t) \Big) &= r_I - \left(\sum_{i \in I} \sum_{j \in J_A(I)} + \sum_{i \in I} \sum_{j \in J_A(I)'} + \sum_{i \in I'} \sum_{j \in J_A(I)} - \sum_{i \in I'} \sum_{j \in J_A(I)} \right) a_{ij}(t) \\ &= r_I - s_{J_A(I)} - a_{I \times J_A(I)'}(t) + a_{I' \times J_A(I)}(t) \\ &\ge r_I - s_{J_A(I)} - 0 + 0. \end{aligned}$$

The last line uses $a_{I \times J_A(I)'}(t) = 0$, induced by $a_{I \times J_A(I)'} = 0$. It also employs the estimate $a_{I' \times J_A(I)}(t) \ge 0$, where equality holds if and only if condition (ii) applies. □

We are now in a position to establish our main result, a formula for the minimum L_1-error. It relies on the marginal vectors r and s, and on the zero pattern of A.

Theorem 3 (Limit). *The limit of the L_1-error function is given by*

$$\lim_{t\to\infty} f(t) = \max_{I\subseteq\{1,\ldots,k\}} \Big(r_I - s_{J_A(I)}\Big).$$

Proof. I. We set $\lambda = \lim_{t\to\infty} f(t)$. Lemma 4 yields the inequality $\lambda \geq \max_{I\subseteq\{1,\ldots,k\}} \big(r_I - s_{J_A(I)}\big)$. With the limit $B = \lim_{n\to\infty} A(t_n)$ of some convergent subsequence along even steps t_n, we prove that its set of underweighted rows, $U_B = \big\{\, i \leq k \;\big|\; b_{i+} < r_i \,\big\}$, achieves equality, $\lambda = r_{U_B} - s_{J_A(U_B)}$. The case $U_B = \emptyset$ is simple. Rows are matching, as are columns. Hence B is a limit biproportional fit and the L_1-error vanishes, $\lambda = 0 = r_\emptyset - s_\emptyset$.

II. The case $U_B \neq \emptyset$ is more complex. With all columns fitted, the L_1-error in B originates exclusively from its rows and, by continuity, coincides with the minimum L_1-error,

$$\lambda = \lim_{n\to\infty} f(t_n) = \frac{1}{2}\sum_{i\leq k}\Big|\lim_{n\to\infty} a_{i+}(t_n) - r_i\Big| = \frac{1}{2}\sum_{i\leq k}\Big|b_{i+} - r_i\Big| > 0.$$

As in Lemma 3 we decompose B into its connected components $B^{(m)}$, $m = 1,\ldots,K$. We have $b_{i+} = \frac{s_{J_m}}{r_{I_m}} r_i < r_i$ if and only if $r_{I_m} > s_{J_m}$. The underweighted rows are $U_B = \bigcup_{m\,:\, r_{I_m} > s_{J_m}} I_m$. With $L = \bigcup_{m\,:\, r_{I_m} > s_{J_m}} J_m$ the corresponding set of columns, we get

$$\lambda = \sum_{i\in U_B}\Big(r_i - b_{i+}\Big) = \sum_{m\,:\, r_{I_m} > s_{J_m}}\;\sum_{i\in I_m}\left(r_i - \frac{s_{J_m}}{r_{I_m}} r_i\right) = r_{U_B} - s_L.$$

III. It remains to show that L coincides with the set $J_A(U_B)$ of columns connected in A to U_B. The inclusion $L \subseteq J_A(U_B)$ is immediate. All columns in L match their marginals and contain a positive entry in some row of U_B, as the rows not in U_B cannot contribute.

The complementary inclusion $L' \subseteq J_A(U_B)'$ needs proof. It means $a_{pq} = 0$, for every $p \in I_m \subseteq U_B$ and $q \in J_M \subseteq L'$. By Lemma 3, the fitting of rows equips the columns in L with divisors $r_{I_m}/s_{I_m} > 1$, whence they are overweighted in $B(1)$. The complement L' then comprises the columns that are matching or underweighted, $r_{I_M} \leq s_{J_M}$.

We claim that convergence of the incremental row divisors $\rho_p(t_n + z)$ to $\frac{s_{J_m}}{r_{I_m}} < 1$ is uniform over all future even steps z, as n tends to infinity. To this end we set $\widetilde{r}_i = \frac{s_{J_m}}{r_{I_m}} r_i$, $\widetilde{\rho}_i(t) = \frac{r_{I_m}}{s_{J_m}} \rho_i(t)$, and $\widetilde{\sigma}_j(t+1) = \frac{s_{J_m}}{r_{I_m}} \sigma_j(t+1)$, for $i \in I_m$ and $j \in J_m$ and $m = 1,\ldots,K$. Then $\widetilde{\rho}_i(t)$ and $\widetilde{\sigma}_j(t+1)$ are the incremental divisors for the fitting of A to $\widetilde{r} = (\widetilde{r}_1,\ldots,\widetilde{r}_k)$ and to s. By Lemma 2(i), $\widetilde{\rho}_{\min}(t)$ and $\widetilde{\rho}_{\max}(t)$ are monotone. Now when $\widetilde{\rho}_p(t_n+z) \geq 1$, then $0 \leq \widetilde{\rho}_p(t_n+z) - 1 \leq \widetilde{\rho}_{\max}(t_n+z) - 1 \leq \widetilde{\rho}_{\max}(t_n) - 1$. Or when $\widetilde{\rho}_p(t_n+z) \leq 1$, then $0 \leq 1 - \widetilde{\rho}_p(t_n+z) \leq 1 - \widetilde{\rho}_{\min}(t_n+z) \leq 1 - \widetilde{\rho}_{\min}(t_n)$. Together this yields

$$\lim_{n\to\infty}\left(\sup_{z=0,2,4,\ldots}\Big|\widetilde{\rho}_p(t_n+z) - 1\Big|\right) \leq \lim_{n\to\infty}\Big(\widetilde{\rho}_{\max}(t_n) - \widetilde{\rho}_{\min}(t_n)\Big) = 1 - 1 = 0.$$

A similar uniformity statement holds for the convergence of the incremental column divisors $\sigma_q(t_n+z+1)$ to $\frac{r_{I_M}}{s_{J_M}} \le 1$. For all $\epsilon > 0$ with $\frac{s_{J_m}}{r_{I_m}} < 1-\epsilon$, there is thus some n_0 such that for all $n \ge n_0$ and for all even z we have $\rho_p(t_n+z) < 1-\epsilon$, and $\sigma_q(t_n+z+1) < 1+\epsilon$. From $a_{pq}(t_n) \le h$ and $x_p(t_n)y_q(t_n) = x_p(t_{n_0})y_q(t_{n_0}) \prod_{z=0}^{t_n-t_{n_0}-2} \rho_p(t_{n_0}+z)\sigma_q(t_{n_0}+z+1)$ we finally get

$$a_{pq} \le h\, x_p(t_{n_0})y_q(t_{n_0}) \lim_{n\to\infty} \left(1-\epsilon^2\right)^{\frac{t_n-t_{n_0}}{2}} = 0.$$

This establishes $L = J_A(U_B)$, and completes the proof. □

6. Convergence of the IPF Procedure. We now exploit the formula for the minimum L_1-error to characterize convergence of the IPF procedure. Theorem 4 handles general limit biproportional fits, Theorem 5 emphasizes those that are disconnected, and Theorem 6 turns to limit biproportional fits that are connected and hence direct. As in Section 5, we assume the components sums of the marginals to be the same, $r_+ = s_+ = h$.

THEOREM 4 (CONVERGENCE). *The following statements are equivalent:*

(1) *The IPF procedure converges.*

(2) *There exists a nonnegative $k \times \ell$ matrix $C = ((c_{ij}))$ inheriting all zeros from A and matching the prespecified marginals, that is, $a_{ij} = 0 \Rightarrow c_{ij} = 0$, and $c_{i+} = r_i$ and $c_{+j} = s_j$, for all rows $i \le k$ and for all columns $j \le \ell$.*

(3) *The partial sums of the marginals satisfy $r_I \le s_{J_A(I)}$ for all $I \subseteq \{1,\dots,k\}$.*

(4) *The L_1-error function converges to zero.*

PROOF. (1) ⇒ (2). If the IPF procedure converges, then $B = \lim_{t\to\infty} A(t)$ clearly qualifies for a matrix C as stipulated in (2).

(2) ⇒ (3). Since all zeros in A are forced upon C we have $c_{I\times J_A(I)'} = 0$, for all $I \subseteq \{1,\dots,k\}$. Nonnegativity of C then implies $r_I = c_{I\times J_A(I)} \le c_{I\times J_A(I)} + c_{I'\times J_A(I)} = s_{J_A(I)}$.

(3) ⇒ (4). From (3) we get $\max_{I\subseteq\{1,\dots,k\}} \left(r_I - s_{J_A(I)}\right) \le 0$, while for $I = \emptyset$ we have $r_\emptyset - s_\emptyset = 0$. Hence Theorem 3 yields $\lim_{t\to\infty} f(t) = 0$.

(4) ⇒ (1). Let B be any accumulation point along a subsequence $A(t_n)$, $n \ge 1$. The L_1-error of B is equal to $\lim_{n\to\infty} f(t_n) = 0$. Hence, with rows and columns fitted, B is a limit biproportional fit. Theorem 1 implies that the sequence $A(t)$, $t \ge 1$, has just a single accumulation point, namely B, and hence converges to B. □

The theorem applies to the situation of Lemma 3, where an accumulation point B is analyzed through its connected components $B^{(m)}$. Given this decomposition, let the *B-adjusted weight matrix* $\widetilde{A}$ punch out the $I_m \times J_m$ submatrices $A^{(m)}$ of A,

$$B = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_K \end{array} \begin{array}{c} \begin{array}{cccc} J_1 & J_2 & \cdots & J_K \end{array} \\ \begin{pmatrix} B^{(1)} & 0 & \cdots & 0 \\ 0 & B^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B^{(K)} \end{pmatrix} \end{array}, \qquad \widetilde{A} = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_K \end{array} \begin{array}{c} \begin{array}{cccc} J_1 & J_2 & \cdots & J_K \end{array} \\ \begin{pmatrix} A^{(1)} & 0 & \cdots & 0 \\ 0 & A^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A^{(K)} \end{pmatrix} \end{array}.$$

We define the *B-adjusted row marginals* $\widetilde{r} = (\widetilde{r}_1, \ldots, \widetilde{r}_k)$ through $\widetilde{r}_i = \frac{s_{J_m}}{r_{I_m}} r_i$, for all $i \in I_m$ and $m = 1, \ldots, K$. They have the same component sum as have the column marginals s. Thus B is the limit biproportional fit, of the original weight matrix A to the B-adjusted row marginals $\widetilde{r}$ and to the original column marginals s. If the IPF procedure is run to fit A to $\widetilde{r}$ and to s, then it converges, by Theorem 4(2), and its limit is B, by Theorem 1. Moreover, if the IPF procedure is employed to fit the B-adjusted weight matrix $\widetilde{A}$ to $\widetilde{r}$ and to s, then it also converges to B, by Theorem 4(2). Now B is the direct biproportional fit, of $\widetilde{A}$ to $\widetilde{r}$ and to s, by Theorem 2.*

In the present section we concentrate on the situations where a limit biproportional fit B, of A to r and to s, exists. In these situations the limit B achieves the prespecified marginals, and there is no need to consider any further adjustments of the row marginals. However, the adjusted weight matrix $\widetilde{A}$ turns out to be quite helpful to understand the disconnectness structure of B.

A cell (i,j) is called *fading* when $a_{ij} > 0$ and $b_{ij} = 0$. An off-diagonal block $I_m \times J_p$ is said to be *fading* when $a_{I_m \times J_p} > 0$ and $b_{I_m \times J_p} = 0$. Our next theorem states that, if B is a limit biproportional fit of A and B is disconnected, then the positions of the fading cells within the matrix A cannot be arbitrary, but satisfy certain specific restrictions. In order to give a formal description of such restrictions, we introduce some graph-theoretical definitions and notation.

Let us associate with the nonnegative matrix A the *positive graph* G_A whose vertices are the positive entries of A and where two positive entries are joined by a *horizontal* edge when they belong to the same row, while they are joined by a *vertical* edge when they belong to the same column. Notice that A is connected if and only if its positive graph

* The first IPF procedure generates the scaled matrices $A(t)$, with $I_m \times J_m$ submatrices $(A(t))^{(m)}$. The second IPF procedure generates the scaled matrices $(A^{(m)})(t)$. These are two distinct entities, for which Bacharach (1970, page 53) uses the same notation A^{2t}_{kk} rather than distinguishing between $(A^{2t})_{kk}$ and $(A_{kk})^{2t}$. Bacharach's "proof by notation" is inconclusive, or so it would seem to us.

G_A is connected, that is, if any two of its vertices are connected by a path. Such paths can be always assumed to be alternating, that is, horizontal and vertical edges alternate along it. The *length* of a path is the number of edges along it. A path or a cycle is said to be *even* when its length is even. A *nasty cycle* is defined to be any even alternating cycle of G_A whose fading cells are at even distance from each other along the cycle.

Let us further associate with A and B the directed graph $\mathcal{G}_{AB}$ whose vertices are the connected components of B and where there is an arc from $B^{(p)}$ to $B^{(m)}$ when the off-diagonal block $I_m \times J_p$ is fading, $a_{I_m \times J_p} > 0$ and $b_{I_m \times J_p} = 0$. The directed graph $\mathcal{G}_{AB}$ is called the *fade digraph* associated with A and B. Notice that the two blocks $B^{(p)}$ and $B^{(m)}$ are uniquely associated with the off-diagonal block $I_m \times J_p$. When B is connected, $\mathcal{G}_{AB}$ is the trivial graph with one vertex.

Finally, let π be a permutation of the rows of B, and τ be a permutation of the columns. The pair (π, τ) is called *diagonal* provided there exists a permutation γ of $\{1, \dots, K\}$ such that $\pi(I_m) = I_{\gamma(m)}$ and $\tau(J_m) = J_{\gamma(m)}$, for all $m = 1, \dots, K$.

THEOREM 5 (FADING BLOCKS). *The following statements are equivalent:*

(1) *B is a limit biproportional fit of A.*

(2) *In the positive graph G_A there is no nasty cycle.*

(3) *The fade digraph $\mathcal{G}_{AB}$ has no circuit.*

(4) *There is a diagonal pair of row and column permutations that makes the matrix A lower block-diagonal.*

PROOF. Let us firstly dispose of the case when B is connected. Then from Theorem 2 we know that B is a direct biproportional fit, so (1) holds true. On the other hand, for $K = 1$ the conditions (2) and (3) trivially hold. So, in the sequel of the proof we may focus on the case when B is disconnected, that is, $K \geq 2$.

(1) $\Rightarrow$ (2). Let B be a limit biproportional fit. Suppose that a nasty cycle (i_1, j_1), $(i_2, j_1), (i_2, j_2), \dots, (i_q, j_q), (i_1, j_q)$ exists in G_A, with a fading initial cell (i_1, j_1). We get

$$\frac{\prod_{p \leq q} a_{i_p j_{p+1}}}{\prod_{p \leq q} a_{i_p j_p}} = \frac{\prod_{p \leq q} a_{i_p j_{p+1}}(t)}{\prod_{p \leq q} a_{i_p j_p}(t)} = \frac{\prod_{p \leq q} b_{i_p j_{p+1}}}{\prod_{p \leq q} b_{i_p j_p}}, \tag{$*$}$$

(where we have set $j_{q+1} = j_1$), since the IPF procedure preserves the initial ratio along all steps t through to the limit. Fading cells contribute zeros only to the numerators. But $(*)$ is a contradiction, as the initial ratio is positive while the limiting ratio is zero.

(2) $\Rightarrow$ (3). Within any connected component $B^{(m)}$, given any row i and any column j, there is always an even alternating path in G_A connecting some entry in row i with some entry in column j. We may then do surgery on the cycle to bypass unwanted portions. Thus we may assume that if the cycle enters and exits some block, it does so just once. Along the cycle, since all fading cells are at even distance from each other, the subpath connecting two consecutive diagonal blocks $B^{(p)}$ and $B^{(m)}$ must consist of a single entry, which necessarily belongs to the off-diagonal block $I_m \times J_p$. Then each fading cell of the cycle gives rise to an arc in $\mathcal{G}_{AB}$. Hence the nasty cycle in G_A induces a circuit in $\mathcal{G}_{AB}$.

(3) $\Rightarrow$ (4). Let the diagonal blocks of B be numbered according to a topological sorting of $\mathcal{G}_{AB}$ (a numbering of the vertices such that each arc is directed from a lower-numbered vertex to a higher-numbered vertex)—always existing in an acyclic digraph. Then all arcs go from some component $B^{(p)}$ to some $B^{(m)}$ with $m > p$. This implies that the connecting off-diagonal fading block $I_m \times J_p$ is located below the diagonal.

(4) $\Rightarrow$ (1). Each component $B^{(m)}$, being the direct biproportional fit of $A^{(m)}$, comes with appropriate divisors u_i and v_j. They give rise to cumulative divisors for the grand problem, $x_i(t) = u_i t^{-m}$ and $y_j(t) = v_j t^p$, for all $i \in I_m$ and $j \in J_p$ and $m, p = 1, \ldots, K$. We now get $a_{ij}(t) = b_{ij}$ if (i, j) belongs to a connected component $(m = p)$. If (i, j) belongs to an off-diagonal block $I_m \times J_p$ with $m > p$, we obtain $a_{ij}(t) = b_{ij} t^{-(m-p)} \to 0$. □

THEOREM 6 (DIRECTNESS). *Suppose A is connected and a limit biproportional fit B exists. Then the following statements are equivalent:*

(1) *B is a direct biproportional fit.*

(2) *There exists a nonnegative $k \times \ell$ matrix $C = ((c_{ij}))$ having the same zeros as A and matching the prespecified marginals, that is, $a_{ij} = 0 \Leftrightarrow c_{ij} = 0$, and $c_{i+} = r_i$ and $c_{+j} = s_j$, for all rows $i \leq k$ and for all columns $j \leq \ell$.*

(3) *The partial sums of the marginals satisfy $r_I < s_{J_A(I)}$ for all row subsets I that are nonempty and proper, $\emptyset \subsetneq I \subsetneq \{1, \ldots, k\}$.*

PROOF. (1) $\Rightarrow$ (2). If B is a direct biproportional fit, then it clearly qualifies for a matrix C as stipulated in (2).

(2) $\Rightarrow$ (3). Let I be a nonempty and proper row subset. The definition of $J_A(I)$ means $a_{I \times J_A(I)'} = 0$ which, in turn, forces $c_{I \times J_A(I)'} = 0$. This implies $c_{I' \times J_A(I)} > 0$, as otherwise C would be disconnected and so would be A, contrary to assumption. We get $r_I = c_{I \times J_A(I)} < c_{I \times J_A(I)} + c_{I' \times J_A(I)} = s_{J_A(I)}$.

(3) $\Rightarrow$ (1). I. We first observe that if the strict inequalities in (3) are satisfied, then so are the inequalities in Theorem 4(3), since the cases $I = \emptyset$ and $I = \{1, \dots, k\}$ hold with equality. By Theorem 4, B is a limit biproportional fit. The proof goes on by contraposition, assuming the limit biproportional fit B not to be direct. Since A is connected, Theorem 2 says that B is disconnected. Hence B has $K \geq 2$ connected components.

II. We next assume that for every $m \in \{1, \dots, K\}$ there is a $p(m) \neq m$ such that the block $I_m \times J_{p(m)}$ is fading. Under this assumption we carry out a scanning process, as follows. We set $m_1 = 1$ and declare block $I_{m_1} \times J_{m_1}$ scanned. With successor $m_2 = p(m_1)$ we pass the fading block $I_{m_1} \times J_{m_2}$ and scan block $I_{m_2} \times J_{m_2}$. Setting $m_3 = p(m_2)$ we transit $I_{m_2} \times J_{m_3}$ to scan $I_{m_3} \times J_{m_3}$, etc. The scanning process terminates with m_F as soon as its successor, $m_S = p(m_F)$ say, is found scanned already. There are just K components, whence $F \leq K$. Discarding the initial section $m_1, \dots, m_{S-1}$ we retain the subsequent path, from starting index m_S up to finishing index m_F. The path consists of $q = F - S + 1 \in \{2, \dots, K\}$ vertices, which we re-label $m_1, \dots, m_q$.

The path gives rise to a circuit with $2q$ vertices in the fade digraph $\mathcal{G}_{AB}$, alternating between on-diagonal blocks $I_{m_p} \times J_{m_p}$ and fading blocks $I_{m_p} \times J_{m_{p+1}}$, for $p = 1, \dots, q$ (setting $m_{q+1} = m_1$). By Theorem 5, the existence of such circuit contradicts the fact that B is a limit biproportional fit. Hence the assumption that every $m \in \{1, \dots, K\}$ comes with a fading block $I_m \times J_{p(m)}$ leads to a contradiction and cannot materialize.

III. The complementary assumption says that for some $m \in \{1, \dots, K\}$ we have $a_{I_m \times J_m'} = 0$. Then J_m are the columns connected in A to I_m, $J_m = J_A(I_m)$. But B fits its marginals, whence $r_{I_m} = s_{J_m}$. We get $r_{I_m} = s_{J_A(I_m)}$, an identity negating (3). □

In the case when all weights in the matrix A are positive, every nonempty and proper row subset I has $J_A(I) = \{1, \dots, \ell\}$ and $r_I - s_{J_A(I)} = r_I - h < 0$. Theorem 4(3) states that the IPF procedure converges to the limit biproportional fit. Theorem 6(3) adds that the fit is direct, in this case.

Acknowledgments. We are grateful to Giles Auchmuty, Norman R. Draper, and Ludger Rüschendorf for valuable remarks on an earlier version of this paper. FP would like to acknowledge the hospitality of the Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza Università di Roma, and support from the Deutsche Forschungsgemeinschaft, during a sabbatical visit 2008/9.

REFERENCES

Bacharach, M. (1965). Estimating nonnegative matrices from marginal data. *International Economic Review (Osaka)* **6** 294–310.

Bacharach, M. (1970). *Biproportional Matrices & Input-Output Change.* Cambridge University Press, Cambridge UK.

Balinski, M.L. and Demange, G. (1989). Algorithms for proportional matrices in reals and integers. *Mathematical Programming* **45** 193–210.

Balinski, M.L. and Pukelsheim, F. (2006). Matrices and politics. In *Festschrift for Tarmo Pukkila on His 60th Birthday* (Eds. E. Liski, J. Isotalo, S. Puntanen, G.P.H. Styan) 233–242. Department of Mathematics, Statistics, and Philosophy, University of Tampere.

Balinski, M.L. and Rachev, S.T. (1997). Rounding proportions: Methods of rounding. *Mathematical Scientist* **22** 1–26.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge MA.

Bregman, L.M. (1967). Proof of the convergence of Sheleikhovskii's method for a problem with transportation constraints. *USSR Computational Mathematics and Mathematical Physics* **7(1)** 191–204.

Brown, D.T. (1959). A note on approximations to discrete probability distributions. *Information and Control* **2** 386-392

Brualdi, R.A., Parter, S.V. and Schneider, H. (1966). The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *Journal of Mathematical Analysis and Applications* **16** 31–50.

Csiszár, I. (1975). *I*-Divergence geometry of probability distributions and minimization problems. *Annals of Probability* **3** 146–158.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11** 427–444.

Fagan, J.T. and Greenberg, B. (1987). Making tables additive in the presence of zeros. *American Journal of Mathematical and Management Sciences* **7** 359-383.

Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics* **41** 907–917.

Fienberg, S.E. and Meyer, M.M. (2006). Iterative proportional fitting. *Encyclopedia of Statistical Sciences* **6** 3723–3726.

Gaffke, N. and Pukelsheim, F. (2008). Vector and matrix apportionment problems and separable convex integer optimization. *Mathematical Methods of Operations Research* **67** 133–159.

Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

Joas, B. (2005). *A graph theoretic solvability check for biproportional multiplier methods.* Thesis, Institut für Mathematik, Universität Augsburg.

Kalantari, B., Lari, I., Ricca, F. and Simeone, B. (2008). On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. *Mathematical Programming Series A* **112** 371–401.

Kullback, S. (1966). An information-theoretic derivation of certain limit relations for a stationary Markov chain. *SIAM Journal on Control* **4** 454–459.

Kullback, S. (1968). Probability densities with given marginals. *Annals of Mathematical Statistics* **39** 1236–1243.

Macgill, S.M. (1977). Theoretical properties of biproportional matrix adjustments. *Environment and Planning A* **9** 687–701.

Maier, S. and Pukelsheim, F. (2007). *Bazi: A Free Computer Program for Proportional Representation Apportionment.* Preprint 42, Institut für Mathematik, Universität Augsburg.

Maier, S., Zachariassen, P. and Zachariasen, M. (2009). Divisor-based biproportional apportionment in electoral systems: A real-life benchmark study. *Management Science*, forthcoming.

MARSHALL, A.W. and OLKIN, I. (1968). Scaling of matrices to achieve specified row and column sums. *Numerische Mathematik* **12** 83–90.

PENNISI, A. (2006). The Italian bug: A flawed procedure for bi-proportional seat allocation. In *Mathematics and Democracy—Recent Advances in Voting Systems and Collective Choice* (Eds. B. Simeone, F. Pukelsheim) 151–166. Springer-Verlag, Berlin.

PRETZEL, O. (1980). Convergence of the iterative scaling procedure for non-negative matrices. *Journal of the London Mathematical Society* **21** 379–384.

PUKELSHEIM, F. (1998). Rounding tables on my bicycle. *Chance* **11** 57–58.

PUKELSHEIM, F. (2004). BAZI—A Java program for proportional representation. *Oberwolfach Reports* **1** 735–737.

PUKELSHEIM, F. and SCHUHMACHER, C. (2004). Das neue Zürcher Zuteilungsverfahren für Parlamentswahlen. *Aktuelle Juristische Praxis—Pratique Juridique Actuelle* **13** 505–522.

RACHEV, S.T. and RÜSCHENDORF, L. (1998). *Mass Transportation Problems. Volume I: Theory. Volume II: Applications.* Springer-Verlag, New York.

RAMÍREZ, V., PUKELSHEIM, F., PALOMARES, A. and MARTÍNEZ, J. (2008). The bi-proportional method applied to the Spanish Congress. *Mathematical and Computer Modelling* **48** 1461–1467.

ROTE, G. and ZACHARIASEN, M. (2007). Matrix scaling by network flow. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Proceedings in Applied Mathematics* **125** 848–854.

RÜSCHENDORF, L. (1995). Convergence of the iterative proportional fitting procedure. *Annals of Statistics* **23** 1160–1174.

SCHNEIDER, M.H. (1990). Matrix scaling, entropy minimization, and conjugate duality (II): The dual problem. *Mathematical Programming* **48** 103–124.

SINKHORN, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics* **35** 877–879.

SINKHORN, R. (1966). A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics* **18** 303–306.

SINKHORN, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly* **74** 403–405.

SINKHORN, R. (1972). Continuous dependence on A in the D_1AD_2 theorems. *Proceedings of the American Mathematical Society* **32** 395–398.

SINKHORN, R. (1974). Diagonal equivalence to matrices with prescribed row and column sums, II. *Proceedings of the American Mathematical Society* **45** 195–198.

SINKHORN, R. and KNOPP, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics* **21** 343–348.

SPEED, T.P. (2005). Iterative proportional fitting. *Encyclopedia of Biostatistics* **7** 2646–2650.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics* **13** 166–178.

WAINER, H. (1998). Visual revelations: Rounding tables. *Chance* **11** 46–50.

ZACHARIASSEN, P. and ZACHARIASEN, M. (2006). A comparison of electoral formulae for the Farœse Parliament. In *Mathematics and Democracy—Recent Advances in Voting Systems and Collective Choice* (Eds. B. Simeone, F. Pukelsheim) 235–252. Springer-Verlag, Berlin.

Friedrich Pukelsheim
Lehrstuhl für Stochastik
und ihre Anwendungen
Universität Augsburg
D-86135 Augsburg, Germany
Pukelsheim@Math.Uni-Augsburg.De

Bruno Simeone
Dipartimento di Statistica,
Probabilità e Statistiche Applicate
Sapienza Università di Roma
I-00185 Roma, Italy
Bruno.Simeone@UniRoma1.It