An EM Approach to OD Matrix Estimation

Robert J. Vanderbei * James Iannone

Program in Statistics & Operations Research
Princeton University
Princeton, NJ 08544

March 29, 1994

Technical Report SOR-94-04

Abstract

Consider a "black box" having I input channels and J output channels. Each arrival on an input channel gets routed through the black box and appears on an output channel. The system is monitored for a fixed time period and a record is made of the number of arrivals on each input channel and the number of departures on each output channel. The OD (origination-destination) matrix estimation problem is to estimate, for each i and j, the number of arrivals on channel i that depart on channel j. We introduce a Poisson stochastic model and employ the EM algorithm to produce high likelihood estimates. In the case of estimation based on observations over a single time-period, we analyze in detail the fixed points of the EM algorithm showing that every vertex of a certain polytope of feasible matrices is a fixed point and identifying a specific interior fixed point which is a saddle point for the likelihood function.

*Research supported by AFOSR through grant AFOSR-91-0359.

Key words: OD matrix estimation, EM algorithm, maximum likelihood estimation.

AMS 1991 Subject Classification: Primary 62H12 Secondary 90B06, 68R05

1 Introduction.

In this paper, we study systems having a set I of input channels and a set J of output channels with the property that each arrival on an input channel gets routed through the system and exits on one of the output channels. For each i and j, we wish to estimate the number of arrivals that entered on i and departed on j. However, such detailed observation is assumed to be impossible (or at least not economical). Instead, the system is monitored for a fixed time period and a record is made of the number of arrivals on each input channel and the number of departures on each output channel but it is not known how the inputs and outputs match up. Based on the observed data, we seek an estimate for the desired *origination-destination (OD) matrix* defined as that matrix whose i, j-th element ($i \in I$, $j \in J$) is the expected number of arrivals on i that depart on j.

Of course, if we were to assume that each arrival instantaneously appears on a departure channel and that there are no simultaneous arrivals, then we could simply set the observation period so small that with overwhelming probability we would see at most one arrival/departure. In this case, there would be no question about the route. Several of these very short observation periods could then be aggregated to form a detailed picture over a longer time horizon. But, in reality, there is typically a delay in the system so that an arrival does not immediately appear on its output channel and this delay is usually of sufficient duration that other arrivals are likely to enter the system during this time. The existence of this delay implies that we should watch the system for an extended period of time so that most of the counted arrivals match up with counted departures. Thus, there is a trade-off between the desire to observe over a short time interval to glean the most information and the need to observe over a long time interval to minimize boundary effects. In this paper, we finesse this issue by assuming on the one hand that there is no system delay but on the other hand that the system is observed over a time interval for which it is likely that there will be several arrivals/departures. For example, consider the problem of modeling traffic flows in a single intersection. In this case, one time interval would be a single cycle through the sequence of lights. The number of cars arriving to and departing from each direction then would typically be less than about twenty. For a comprehensive account of this traffic management problem see [4].

Suppose that an observation time-interval has been determined. Furthermore, we assume that data will be collected over multiple time intervals. One can envision two possible objectives: either

- to collect data for a fixed number of time intervals and then produce an estimate at the end, or
- to develop a running estimate of the OD matrix as the system evolves.

In this paper, we focus entirely on the second objective. To produce a running estimate, it is natural to consider iterative procedures in which a current estimate, say $\tilde{\Lambda}^{(t)}$, is updated to a new estimate $\tilde{\Lambda}^{(t+1)}$ using only the statistics observed over the current time interval. In fact, probably the simplest and most reasonable procedure would be to use a time series,

$$\tilde{\Lambda}^{(t+1)} = p\tilde{\Lambda}^{(t)} + q\tilde{\Lambda}^{(t,cur)},$$

where p and q are nonnegative real numbers summing to one and $\tilde{\Lambda}^{(t,cur)}$ denotes an estimate for the OD matrix computed using only the current observations. In this paper, we use the EM algorithm to produce the estimate $\tilde{\Lambda}^{(t,cur)}$. In particular, we derive the formula for this algorithm, study its properties, and discuss efficient mechanisms for implementing the algorithm. It should be noted from the start that one cannot expect any estimator of the OD matrix computed based on one obsevation period to produce a very good estimator since this is a problem of estimating $|I| \times |J|$ parameters using only $|I| + |J| - 1$ observed data points. Nonetheless, one hopes that the time-series approach will eventually produce a good estimate. Numerical studies confirming this hope can be found in [2].

Henceforth, we shall focus on only one observation interval. We assume that, for each $i \in I$ and $j \in J$, the number Y_{ij} of arrivals on channel i that depart on channel j is a Poisson random variable with (unknown) parameter λ_{ij}. Furthermore, we assume that these random variables are independent of each other. For each $i \in I$, we let

$$X_i = \sum_{j \in J} Y_{ij}$$

denote the total number of arrivals on channel i and, for each $j \in J$, we let

$$Z_j = \sum_{i \in I} Y_{ij}$$

denote the number of departures on channel j. Since the sum of independent Poisson random variables is also Poisson, it follows that the vector

$$X = (X_i : i \in I)$$

is a collection of independent Poisson random variables. Similarly, the vector

$$Z = (Z_j : j \in J)$$

is a collection of independent Poisson random variables. However, the random variables in X are not independent of the random variables in Z.

The problem is to estimate the matrix of unknown parameters

$$\Lambda = (\lambda_{ij} : i \in I, j \in J)$$

based on observations of X and Z. In this paper, we look for a maximum likelihood estimator for this matrix. Indeed, given a vector

$$m = (m_i : i \in I)$$

of arrival observations and a vector

$$n = (n_j : j \in J)$$

of departure observations, the likelihood of Λ is given by

$$\begin{aligned} L(\Lambda) &= \mathbf{P}(X = m, Z = n) \\ &= \sum_{R \in \mathcal{R}(m,n)} \mathbf{P}(Y = R) \\ &= \sum_{R \in \mathcal{R}(m,n)} \prod_{i,j} e^{-\lambda_{ij}} \frac{\lambda_{ij}^{r_{ij}}}{r_{ij}!}, \end{aligned} \tag{1.1}$$

where $\mathcal{R}(m,n)$ denotes the set of matrices $R = (r_{ij} : i \in I, j \in J)$ having nonnegative integer elements and having the correct row and column sums:

$$\begin{aligned} \sum_j r_{ij} &= m_i \quad \text{for each } i \in I \\ \sum_i r_{ij} &= n_j \quad \text{for each } j \in J \end{aligned}$$

Our goal is to maximize this likelihood function over all Λ in $\mathbb{R}_+^{I\times J}$. In Section 2, we discuss this optimization problem and the EM algorithm which attempts to find a maximum. In Sections 3 and 4, we study the fixed points of the EM algorithm. Section 5 contains the details of a specific example and finally Section 6 discusses computational issues. We end this section with some definitions that we will need throughout the paper.

1.1 Notations and Definitions

Before continuing, we introduce some economizing notation that will be used throughout this paper. If Λ is a matrix with nonnegative real entries and R is a matrix with nonnegative integer entries, we let Λ^R denote the obvious componentwise product

$$\Lambda^R = \prod_{ij} \lambda_{ij}^{r_{ij}}$$

and we let $R!$ denote the componentwise product of factorials

$$R! = \prod_{ij} r_{ij}!.$$

Also, we let

$$|\Lambda| = \sum_{i,j} \lambda_{ij}$$

and similarly for a vector $m = (m_i : i \in I)$ we let

$$|m| = \sum_i m_i.$$

Throughout this paper m will denote the vector of observed arrivals, n will denote the vector of observed departures and

$$N = |m| = |n|$$

will denote the total number of arrivals/departures.

With these notations, the likelihood function can be written as follows:

$$L(\Lambda) = \sum_{R\in\mathcal{R}} e^{-|\Lambda|}\frac{\Lambda^R}{R!}, \tag{1.2}$$

where we have written $\mathcal{R}$ instead of $\mathcal{R}(m,n)$. We will typically not mention the dependence of $\mathcal{R}$ on m and n unless it is necessary.

Finally, the *support* of a matrix $R = (r_{ij} : i \in I, j \in J)$ is the subset of $I \times J$ on which $r_{ij} \neq 0$.

2 Maximum Likelihood and the EM Algorithm

The EM algorithm [1, 8] is an algorithm for finding high likelihood estimators from incomplete data. It is an iterative algorithm in which a starting estimate Λ^0 is iteratively updated according to a formula

$$\Lambda^{k+1} = f(\Lambda^k),$$

where f is a transformation on $\mathbb{R}_+^{I \times J}$. The transformation f is called the *EM operator.* It is known [1] that the EM operator generates a sequence of estimates that have monotonically increasing likelihood. Hence, even if the algorithm fails to produce a maximum likelihood estimate (and failure is typical), it does always produce an increasing sequence and so we describe the limiting estimate as a high likelihood estimator. In this section, we give two separate derivations of the EM operator.

See [6, 7] for applications of the EM algorithm to similar estimation problems.

2.1 First Derivation

The first derivation is the simplest and has nothing to do with the usual description of how one obtains an EM algorithm. Instead, we simply suppose that the maximum likelihood estimator lies in the interior of $\mathbb{R}_+^{I \times J}$ and differentiate the likelihood function to find its interior critical points. Setting to zero the derivative with respect to λ_{kl}, we see that

$$0 = \frac{\partial L}{\partial \lambda_{kl}} = \sum_{R \in \mathcal{R}} e^{-|\Lambda|} \frac{\Lambda^R}{R!} \left(\frac{r_{kl}}{\lambda_{kl}} - 1 \right). \tag{2.1}$$

Now, for $\lambda_{kl} > 0$, we can multiply through by λ_{kl} to obtain the following equivalent expression:

$$\sum_{R \in \mathcal{R}} e^{-|\Lambda|} \frac{\Lambda^R}{R!} r_{kl} = \lambda_{kl} \sum_{R \in \mathcal{R}} e^{-|\Lambda|} \frac{\Lambda^R}{R!}. \tag{2.2}$$

Any interior matrix Λ that satisfies (2.2) for every $k \in I$, $l \in J$ is a critical point for the likelihood function. Cancelling $\exp(-|\Lambda|)$ from both sides and writing the equations indexed by $k \in I$, $l \in J$ as one matrix equation, we see that interior critical points must satisfy

$$\Lambda = f(\Lambda), \tag{2.3}$$

where f is the transformation on $\mathbb{R}_+^{I \times J}$ defined by

$$f(\Lambda) = \frac{\displaystyle\sum_{R \in \mathcal{R}} \frac{\Lambda^R}{R!} R}{\displaystyle\sum_{R \in \mathcal{R}} \frac{\Lambda^R}{R!}}. \tag{2.4}$$

As we shall see next, this is the EM operator.

2.2 Second Derivation

The EM algorithm is usually defined as follows (see, e.g., [8]). First, compute the maximum likelihood estimator assuming full information (the M-step). Then take the conditional expectation (the E-step) of this maximum likelihood estimator (computed using the current value of Λ) given only the observed information.

For the problem at hand, full information means knowing the values of Y. Hence, the full information likelihood function is

$$\tilde{L}(\Lambda) = \mathbf{P}(Y = R) = e^{-|\Lambda|} \frac{\Lambda^R}{R!}.$$

Since this likelihood function corresponds to independent Poisson random variables, the full information maximum likelihood estimator is easy to compute:

$$\Lambda = R,$$

or, in terms of random variables,

$$\Lambda = Y.$$

Then the E-step of the EM algorithm produces the next iterate:

$$\tilde{\Lambda} \;\; = \;\; \mathbf{E}[Y | X = m, Z = n]$$

$$
\begin{aligned}
&= \frac{\sum_{R \in \mathcal{R}} e^{-|\Lambda|} \frac{\Lambda^R}{R!} R}{\sum_{R \in \mathcal{R}} e^{-|\Lambda|} \frac{\Lambda^R}{R!}} \\
&= f(\Lambda),
\end{aligned}
$$

where f is the EM operator defined by (2.4).

Remark: Note that this derivation of the EM algorithm differs slightly from the usual derivation as given, e.g., in [1] or [8]. In Appendix A, we give a derivation that conforms exactly with the cited references. We include this appendix to make it clear that equation (2.4) is the EM algorithm.

3 Boundary Fixed Points of the EM Operator

In this section, we begin our study of the fixed points of the EM operator f defined by (2.4). First, note that every fixed point must belong to the convex hull of $\mathcal{R}$ since f can be viewed as forming a convex combination of the matrices R in $\mathcal{R}$.

Theorem 1 *Every extreme point of the convex hull of $\mathcal{R}$ is a fixed point of the EM operator.*

The proof of this theorem depends on the following lemma:

Lemma 2 *Given a matrix $\Delta \neq 0$ with integer entries whose row and column sums vanish, there exists a matrix $U \neq 0$ with entries in $\{-1, 0, 1\}$ such that the support of U is contained in the support of Δ, and the row and column sums of U vanish.*

Proof. The matrix Δ represents a bipartite graph as follows. The left-hand nodes correspond to the rows of Δ (i.e., they correspond to the set of input channels I), the right-hand nodes correspond to the columns of Δ (i.e., the output channels J), every positive entry in Δ, say Δ_{ij} corresponds to a link from i to j having flow equal to Δ_{ij}, and finally, every negative Δ_{ij} corresponds to a link from j to i having flow equal to $-\Delta_{ij}$. The fact that Δ has vanishing row and column sums means that there is flow balance at each node of the bipartitie graph. The flow decomposition theorem for bipartite graphs (see e.g. [5], Theorem 3.5) states that this flow can be decomposed as a sum of unit flows. Since $\Delta \neq 0$, this decomposition is nontrivial. The desired matrix U is the matrix corresponding to any of the unit flows in this decompostion. $\square$

Proof of Theorem 1. Let S be an extreme point of the convex hull of $\mathcal{R}$. Since the right-hand side in (2.4) is a convex combination of the R's in $\mathcal{R}$, we see that S is a fixed point of the EM operator if and only if

$$\frac{S^R}{R!} = 0$$

for all $R \in \mathcal{R}\backslash\{S\}$. Suppose there exists an $R \in \mathcal{R}\backslash\{S\}$ such that $S^R/R! \neq 0$. Then, it is easy to see that the support of R is contained in the support of S. Let $\Delta = S - R$. Clearly, Δ is a matrix with integer entries whose row and column sums vanish and whose support is contained in the support of S. Now, from Lemma 2 there exists a nonzero matrix U whose support is contained in the support of Δ and whose row and column sums vanish. Since all nonzero elements of U are ± 1, it follows that $S + U$ and $S - U$ belong to $\mathcal{R}$. Therefore, S is not an extreme point, which is a contradiction. Hence, $S^R/R! = 0$ for all $R \in \mathcal{R} \setminus \{S\}$. $\square$

4 An Interior Fixed Point of the EM Operator

In this section, we exhibit an interior fixed point. In proving that the interior point (given by (4.1)) is a fixed point, we will need a multinomial generalization of the Vandermonde convolution identity. We find this result interesting in its own right and hence we present it first.

Recall that $N = |m| = \sum_i m_i$. We use the standard notation for multinomial coefficients:

$$\left(\begin{array}{c} N \\ m \end{array} \right) = \frac{N!}{\prod_i m_i!}$$

representing the number of ways that N items can be divided into sets such that m_i items are assigned to the i-th set.

Theorem 3 *The following multinomial identity holds:*

$$\sum_{T \in \mathcal{R}(m,n)} \prod_j \left(\begin{array}{c} n_j \\ T_{\cdot j} \end{array} \right) = \left(\begin{array}{c} N \\ m \end{array} \right),$$

where $T_{\cdot j}$ denotes the j-th column of matrix T.

For 2×2 matrices, this theorem reduces exactly to the well-known Vandermonde convolution identity:

$$\sum_k \left(\begin{array}{c} n \\ k \end{array} \right) \left(\begin{array}{c} N-n \\ m-k \end{array} \right) = \left(\begin{array}{c} N \\ m \end{array} \right).$$

Theorem 3 provides what appears to be the most natural generalization of this identity. This generalization has also recently appeared in the context of bi-multivariate hypergeometric distributions [3]. While the general result does not seem to follow in any direct way from the 2×2 result, the proof for the 2×2 case generalizes quite naturally:

Proof. We count by two methods the number of ways that N items can be assigned to groups indexed by I such that the i-th group receives m_i items. Call these groups *row groups*. Introduce another independent set of groups, called *column groups*, indexed by J and suppose that we wish to put n_j items into the j-th column group. Consider an $I \times J$ matrix of groups and suppose that we wish to assign t_{ij} items to the (i,j)-th group. Let T denote the matrix whose elements are t_{ij}. For fixed $j \in J$, the number of ways of assigning t_{1j} items to group 1, t_{2j} items to group 2, and so on, is given by

$$\left(\begin{array}{c} |T_{\cdot j}| \\ T_{\cdot j} \end{array} \right).$$

Since we wish the total number of items assigned to column j to be n_j, it follows that $|T_{\cdot j}| = n_j$. Furthermore, the numbers of ways items can be distributed into distinct columns are independent from one another. Hence, the number of ways of assigning t_{ij} items to group (i,j) as (i,j) ranges over all of $I \times J$ is given by

$$\prod_j \binom{|T_{\cdot j}|}{T_{\cdot j}}.$$

Finally, summing over all T in $\mathcal{R}(m,n)$ yields the number of ways of distributing N items into row groups such that group i gets m_i items. □

Theorem 4 *The interior point*

$$\Lambda = \frac{\sum_{R\in\mathcal{R}} R/R!}{\sum_{R\in\mathcal{R}} 1/R!} \tag{4.1}$$

is a fixed point of the EM operator.

Proof. To show that a matrix Λ is a fixed point of f given by (2.4), it suffices to show that

$$\sum_{R\in\mathcal{R}} \frac{\Lambda^R}{R!}(R-\Lambda) = 0. \tag{4.2}$$

Substituting the specific matrix Λ given by (4.1) into the left-hand side of (4.2) and pulling out common denominators, (4.2) reduces to

$$\frac{1}{\left(\sum_{S\in\mathcal{R}} 1/S!\right)^{N+1}} \sum_{R\in\mathcal{R}} \frac{\left(\sum_{T\in\mathcal{R}} T/T!\right)^R}{R!} \sum_{S\in\mathcal{R}} \frac{1}{S!}(R-S) = 0.$$

Hence, it is necessary and sufficient to check that

$$\sum_{R\in\mathcal{R}} \sum_{S\in\mathcal{R}} \frac{\left(\sum_{T\in\mathcal{R}} T/T!\right)^R}{R!S!}(R-S) = 0.$$

This follows immediately from the following amazing lemma. □

Lemma 5 *The real-valued function g defined on $\mathcal{R}$ by the formula*

$$g(R) = (\sum_{T\in\mathcal{R}} T/T!)^R$$

is constant.

Proof. We say that two matrices R and S in $\mathcal{R}$ are neighbors if they differ by a matrix that has all zero entries except for two 1's and two -1's configured as follows:

$$S - R = \begin{array}{c} \\ i_1 \\ i_2 \\ \\ \end{array} \left[\begin{array}{cc} j_1 & j_2 \\ 1 & -1 \\ -1 & 1 \\ & \end{array} \right] \tag{4.3}$$

It is easy to see that, with this neighbor relation, $\mathcal{R}$ is connected. Hence, to show that g is constant, it suffices to show that $g(R) = g(S)$ for every pair of neighbors R and S in $\mathcal{R}$. From (4.3) it is easy to see that

$$\left(\sum_{T\in\mathcal{R}} \frac{T}{T!}\right)^S = \left(\sum_{T\in\mathcal{R}} \frac{T}{T!}\right)^R \frac{\left(\sum_{T\in\mathcal{R}} \frac{t_{i_1j_1}}{T!}\right)\left(\sum_{T\in\mathcal{R}} \frac{t_{i_2j_2}}{T!}\right)}{\left(\sum_{T\in\mathcal{R}} \frac{t_{i_1j_2}}{T!}\right)\left(\sum_{T\in\mathcal{R}} \frac{t_{i_2j_1}}{T!}\right)}.$$

Thus, g is constant if and only if

$$\left(\sum_{T\in\mathcal{R}} \frac{t_{i_1j_1}}{T!}\right)\left(\sum_{T\in\mathcal{R}} \frac{t_{i_2j_2}}{T!}\right) = \left(\sum_{T\in\mathcal{R}} \frac{t_{i_1j_2}}{T!}\right)\left(\sum_{T\in\mathcal{R}} \frac{t_{i_2j_1}}{T!}\right). \tag{4.4}$$

Now, a simple change of variables shows that

$$\sum_{T\in\mathcal{R}(m,n)} \frac{t_{ij}}{T!} = \sum_{T\in\mathcal{R}(m-e_i,n-e_j)} \frac{1}{T!},$$

where e_i denotes the unit vector having a one in position i. Hence g is constant if and only if, for every choice of $i_1, i_2 \in I$ and every choice of $j_1, j_2 \in J$,

$$\left(\sum_{T\in\mathcal{R}(m-e_{i_1},n-e_{j_1})} \frac{1}{T!}\right)\left(\sum_{T\in\mathcal{R}(m-e_{i_2},n-e_{j_2})} \frac{1}{T!}\right)$$

$$= \left(\sum_{T \in \mathcal{R}(m-e_{i_1}, n-e_{j_2})} \frac{1}{T!} \right) \left(\sum_{T \in \mathcal{R}(m-e_{i_2}, n-e_{j_1})} \frac{1}{T!} \right). \tag{4.5}$$

Finally, Theorem 3 shows that, for any m and n,

$$\sum_{T \in \mathcal{R}(m,n)} \frac{1}{T!} = \binom{N}{m} \frac{1}{\prod_j n_j!} = \frac{N!}{(\prod_i m_i!)(\prod_j n_j!)}, \tag{4.6}$$

which reduces the verification of (4.5) to triviality. □

Corollary 6 (Hindsight) *The expression for the interior point* Λ *given by (4.1) simplifies to the following:*

$$\Lambda = \frac{m \otimes n}{N},$$

i.e., $\lambda_{ij} = m_i n_j / N$.

Proof. From (4.1), the (i,j)-th element of Λ is given by

$$\begin{aligned} \lambda_{ij} &= \frac{\displaystyle\sum_{R \in \mathcal{R}(m,n)} \frac{r_{ij}}{R!}}{\displaystyle\sum_{R \in \mathcal{R}(m,n)} \frac{1}{R!}} \\ &= \frac{\displaystyle\sum_{R \in \mathcal{R}(m-e_i, n-e_j)} \frac{1}{R!}}{\displaystyle\sum_{R \in \mathcal{R}} \frac{1}{R!}}. \end{aligned}$$

Now, (4.6) gives explicit expressions for the numerator and denominator which combine to yield that

$$\lambda_{ij} = \frac{m_i n_j}{N}.$$

□

The preceeding results suggest that perhaps the likelihood function is convex (or concave). However, simple examples show that the likelihood function (1.2) is neither concave nor convex on all of $\mathbb{R}_+^{I\times J}$ but for all examples that we have analyzed it is strictly convex when its domain is restricted to the convex hull of $\mathcal{R}(m,n)$. If such a statement could be proved in general, it would imply that the interior feasible point given in Theorem 4 is the only interior feasible point, that it is a minimum likelihood point relative to the convex hull of $\mathcal{R}(m,n)$, and, in particular, that it has lower likelihood than the extreme point solutions of Theorem 1. However, we have been unable to prove this in general and so we are content with the following partial result.

Theorem 7 *The interior fixed point given in Theorem 4 is not a local maximum of the likelihood function.*

Proof. The likelihood function is given in (1.2) and its first derivatives are given in (2.1). Its second derivatives are

$$\frac{\partial^2 L}{\partial \lambda_{ij}\partial \lambda_{kl}} = \sum_{R\in\mathcal{R}} e^{-|\Lambda|}\frac{\Lambda^R}{R!}\left[\left(\frac{r_{ij}}{\lambda_{ij}}-1\right)\left(\frac{r_{kl}}{\lambda_{kl}}-1\right) - \delta_{(i,j)=(k,l)}\frac{r_{ij}}{\lambda_{ij}^2}\right]$$

and the associated quadratic form is given by

$$\begin{aligned} Q(\xi) &= \sum_{i,j,k,l}\xi_{ij}\frac{\partial^2 L}{\partial\lambda_{ij}\partial\lambda_{kl}}\xi_{kl} \\ &= \sum_{R\in\mathcal{R}} e^{-|\Lambda|}\frac{\Lambda^R}{R!}\left[\left(\sum_{i,j}\left(\frac{r_{ij}}{\lambda_{ij}}-1\right)\xi_{ij}\right)^2 - \sum_{i,j}\frac{r_{ij}}{\lambda_{ij}^2}\xi_{ij}^2\right]. \end{aligned} \tag{4.7}$$

Fixing specific values $i_1, i_2 \in I$ and $j_1, j_2 \in J$ and choosing

$$\xi = [\xi_{ij}] = \begin{array}{c} \\ i_1 \\ i_2 \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} j_1 & j_2 \end{array} \\ \left[\begin{array}{cc} & \\ \zeta & -\zeta \\ -\zeta & \zeta \\ & \end{array}\right] \end{array},$$

the bracketed expression in (4.7) reduces to ζ^2 times the following expression:

$$
\begin{array}{ccccccc}
\dfrac{r_{i_1j_1}(r_{i_1j_1}-1)}{\lambda^2_{i_1j_1}} & + & 2\dfrac{r_{i_1j_1}r_{i_2j_2}}{\lambda_{i_1j_1}\lambda_{i_2j_2}} & - & 2\dfrac{r_{i_1j_1}r_{i_1j_2}}{\lambda_{i_1j_1}\lambda_{i_1j_2}} & - & 2\dfrac{r_{i_1j_1}r_{i_2j_2}}{\lambda_{i_1j_1}\lambda_{i_2j_1}} \\
 & + & \dfrac{r_{i_2j_2}(r_{i_2j_2}-1)}{\lambda^2_{i_2j_2}} & - & 2\dfrac{r_{i_2j_2}r_{i_1j_2}}{\lambda_{i_2j_2}\lambda_{i_1j_2}} & - & 2\dfrac{r_{i_2j_2}r_{i_2j_1}}{\lambda_{i_2j_2}\lambda_{i_2j_1}} \\
 & & & + & \dfrac{r_{i_1j_2}(r_{i_1j_2}-1)}{\lambda^2_{i_1j_2}} & + & 2\dfrac{r_{i_1j_2}r_{i_2j_1}}{\lambda_{i_1j_2}\lambda_{i_2j_1}} \\
 & & & & & + & \dfrac{r_{i_2j_1}(r_{i_2j_1}-1)}{\lambda^2_{i_2j_1}}.
\end{array}
$$

Hence, for this choice of ξ, the quadratic form becomes

$$
\begin{aligned}
Q(\xi) = e^{-|\Lambda|}\zeta^2 \quad [\quad & S(m-2e_{i_1}, n-2e_{j_1}) \\
+ \quad & 2S(m-e_{i_1}-e_{i_2}, n-e_{j_1}-e_{j_2}) \\
- \quad & 2S(m-2e_{i_1}, n-e_{j_1}-e_{j_2}) \\
- \quad & 2S(m-e_{i_1}-e_{i_2}, n-2e_{j_1}) \\
\\
+ \quad & S(m-2e_{i_2}, n-2e_{j_2}) \\
- \quad & 2S(m-e_{i_1}-e_{i_2}, n-2e_{j_2}) \\
- \quad & 2S(m-2e_{i_2}, n-e_{j_1}-e_{j_2}) \\
\\
+ \quad & S(m-2e_{i_1}, n-2e_{j_2}) \\
+ \quad & 2S(m-e_{i_1}-e_{i_2}, n-e_{j_1}-e_{j_2}) \\
\\
+ \quad & S(m-2e_{i_2}, n-2e_{j_1}) \quad],
\end{aligned}
\tag{4.8}
$$

where

$$S(m', n') = \sum_{r \in \mathcal{R}(m',n')} \frac{\Lambda^R}{R!}.$$

Now, using the explicit formula for the interior fixed point given in Corollary 6 together with (4.6), we see that

$$S(m', n') = \frac{\prod_i m_i^{m'_i} \prod_j n_j^{n'_j}}{N^{N'}} \frac{N'!}{\prod_i m'_i! \prod_j n'_j!},$$

where $N' = |m'| = |n'|$. Substituting this formula for $S(m', n')$ into (4.8), we get

$$\begin{aligned} Q(\xi) = e^{-|\Lambda|}\zeta^2 S(m,n) \quad \Bigg[& \frac{(m_{i_1}-1)}{m_{i_1}}\frac{(n_{j_1}-1)}{n_{j_1}} \\ + \; & 2 \\ - \; & 2\frac{(m_{i_1}-1)}{m_{i_1}} \\ - \; & 2\frac{(n_{j_1}-1)}{n_{j_1}} \\ + \; & \frac{(m_{i_2}-1)}{m_{i_2}}\frac{(n_{j_2}-1)}{n_{j_2}} \\ - \; & 2\frac{(n_{j_2}-1)}{n_{j_2}} \\ - \; & 2\frac{(m_{i_2}-1)}{m_{i_2}} \\ + \; & \frac{(m_{i_1}-1)}{m_{i_1}}\frac{(n_{j_2}-1)}{n_{j_2}} \\ + \; & 2 \\ + \; & \frac{(m_{i_2}-1)}{m_{i_2}}\frac{(n_{j_1}-1)}{n_{j_1}} \Bigg], \end{aligned} \qquad (4.9)$$

which finally simplifies to

$$Q(\xi) = e^{-|\Lambda|}\zeta^2 S(m,n)\left(\frac{1}{m_{i_1}} + \frac{1}{m_{i_2}}\right)\left(\frac{1}{n_{j_1}} + \frac{1}{n_{j_2}}\right) \geq 0.$$

Hence, $\Lambda = mn^T/N$ is not a local maximum. □

Remark: In this section, we have spent a great deal of effort to study an interior fixed point which turns out not even to be a local maximum (in fact, as the example of the next section shows, it appears to be the global minimum of the likelihood function restricted to conv($\mathcal{R}$)). Hence, this is a solution one wishes to avoid. Unfortunately, it is easy to stumble upon this solution. Indeed, if one starts the EM algorithm by setting the initial choice of Λ to the matrix of all ones, then the first iteration of the EM algorithm replaces this initial choice with the interior fixed point (4.1). This is precisely how we discovered it!

5 An Example

Suppose that $|I| = |J| = 2$. Then $m = (m_1, m_2)$ and $n = (n_1, n_2)$. In this case, $\mathcal{R}$ has only two extreme points:

$$\begin{bmatrix} m_1 \wedge n_1 & m_1 - m_1 \wedge n_1 \\ n_1 - m_1 \wedge n_1 & m_2 \wedge n_2 \end{bmatrix}$$

and

$$\begin{bmatrix} m_1 - m_1 \wedge n_2 & m_1 \wedge n_2 \\ m_2 \wedge n_1 & n_2 - m_1 \wedge n_2 \end{bmatrix}.$$

Note that the first extreme point is obtained by applying the the northwest-corner method and the second extreme point is obtained by applying the analogous method starting in the northeast corner (note also that various identities, such as $m_1 - m_1 \wedge n_1 + m_2 \wedge n_2 = n_2$, follow from the fact that $m_1 + m_2 = n_1 + n_2$). Theorem 1.1 shows that both of these extreme points are fixed points of the EM algorithm. Hence, we could always select the first one. This results in the following EM estimator for λ_{11}:

$$\hat{\Lambda}_{11} = X_1 \wedge Z_1.$$

This EM estimator is a biased estimator:

$$\begin{aligned} \mathbf{E}\hat{\Lambda}_{11} &= \mathbf{E}X_1 \wedge Z_1 \\ &= \mathbf{E}Y_{11} + \mathbf{E}\left(X_1 \wedge Z_1 - Y_{11}\right) \\ &> \mathbf{E}Y_{11} \\ &= \lambda_{11}. \end{aligned}$$

We make the following conjecture:

Conjecture 8 *If the EM algorithm is initialized with an estimate Λ that is uniformly distributed over the convex hull of $\mathcal{R}$, then the resulting EM estimator (which will be a random convex combination of the extreme points) is unbiased.*

Now suppose that $m = (2, 1)$ and $n = (2, 1)$. In this case the two extreme points are

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

and the convex hull can be parametrized with a single real parameter p:

$$p \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} + (1-p) \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1+p & 1-p \\ 1-p & p \end{bmatrix}, \quad 0 \leq p \leq 1.$$

Hence, the likelihood function can also be expressed as a function of p:

$$\begin{aligned} L(p) &= e^{-3} \left(\frac{(1+p)^2 p}{2} + (1+p)(1-p)^2 \right) \\ &= e^{-3} \left(\frac{3}{2} p^3 - \frac{1}{2} p + 1 \right). \end{aligned}$$

From this explicit formula, it is easy to see that L is convex on $[0, 1]$, has a minimum at $p = 1/3$, a local maximum at $p = 0$ and a global maximum at $p = 1$.

Furthermore, the EM function viewed as a function from $[0, 1]$ to $[0, 1]$ has the following form:

$$f(p) = \frac{p^2 + p}{2p^2 - 3p + 2}.$$

It is easy to see that the EM algorithm will converge to the global maximum if $p > 1/3$, to the other local maximum if $p < 1/3$ and to the minimum if $p = 1/3$. Hence, if the algorithm is started by choosing a value of p uniformly from $[0, 1]$, then with probability $2/3$ the EM algorithm will converge to the global maximum.

As just shown for this example, the likelihood function restricted to the convex hull of $\mathcal{R}$ is a convex function. We have been unable to determine whether or not this holds in general.

6 Computational Issues

Computing the EM function f can be a formidable task since the number of matrices in $\mathcal{R}$ is generally huge. For example, for $m = (3, 2, 3, 4)$, and $n = (3, 5, 1, 3)$, the cardinality of $\mathcal{R}(m, n)$ is 797 and for $m = (8, 3, 3, 5)$ and

$n = (4, 4, 6, 5)$, the cardinality of $\mathcal{R}(m, n)$ is 12527. Hence, we seek a more efficient computational procedure.

Since every extreme point is known to be a fixed point of the EM function (and hence an EM estimator), we could apply the northwest corner rule to find one quickly. However, as indicated in the example of the previous section, this process will generally produce a biased estimator. One would expect to get a better estimator by first randomly permuting the rows and the columns before applying the northwest corner rule. This procedure will yield a random vertex. While this procedure appears attractive from a computational point of view, it too seems to have limitations. Indeed, on the example of the previous section it will produce one of the two extreme points each with probability 1/2. However, if we start the EM algorithm using a randomly generated starting point, there will be a tendency to favor the extreme point that is the global maximum. Hence, even though generating extreme points using a northwest-corner type rule is computationally attractive, one might prefer to actually carry out the EM algorithm and let it converge to a vertex that hopefully has high likelihood. But, as mentioned above, direct calculation of the likelihood function is impossible for problems of any significant size. To diminish this computational problem, we introduce a Markov chain R_k, $k = 0, 1, 2, \ldots$, on the set $\mathcal{R}$ whose limiting distribution is known to be uniform over $\mathcal{R}$. Hence, from the ergodic theorem we get that

$$f(\Lambda) = \lim_{k\to\infty} \frac{\sum_{i=0}^{k} \frac{\Lambda^{R_i}}{R_i!} R_i}{\sum_{i=0}^{k} \frac{\Lambda^{R_i}}{R_i!}}.$$

By simply summing up to a large value of k we arrive at a computational procedure that is reasonably effective (see [2] for more details).

The random walk is defined as follows. Starting from a matrix $R \in \mathcal{R}$, select at random two indices $i_1, i_2 \in I$, two indices $j_1, j_2 \in J$, a random integer ζ between $-r_{i_1 j_1} \wedge r_{i_2 j_2}$ and $r_{i_1 j_2} \wedge r_{i_2 j_1}$ and step to a new point

$$S = R \quad + \quad \begin{array}{c} \\ i_1 \\ i_2 \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} j_1 & j_2 \end{array} \\ \left[\begin{array}{ccc} & & \\ & \zeta & -\zeta \\ & -\zeta & \zeta \\ & & \end{array} \right] \end{array} .$$

It is easy to see that this defines an irreducible Markov chain for which the probability transition matrix is symmetric and hence doubly stochastic. Therefore, the limiting distribution exists and is the uniform distribution on $\mathcal{R}$.

7 Concluding Remarks

We have studied the EM algorithm as it applies to the problem of estimating an OD matrix given its marginals. The operator defining the EM algorithm has many fixed points including every extreme point on the convex hull of feasible matrices. We have found a specific expression for one interior point of the convex hull and proved that this point is not a local maximum.

This algorithm has been implemented and tested on randomly generated data and has performed reasonably well for the trial data sets. Some improvements are made by considering multiple time periods at each iteration and then recursively determining an estimate using this window in time. The accuracy of these results are remarkable considering the lack of data that one has in estimating the OD matrix.

A Third Derivation

The third derivation of the EM operator follows the mathematical derivation given in [8]. We start with the full information likelihood function

$$\tilde{L}(\tilde{\Lambda}) = \mathbf{P}_{\tilde{\Lambda}}(Y = R) = e^{-|\tilde{\Lambda}|}\frac{\tilde{\Lambda}^R}{R!}$$

and compute the conditional expectation of its logarithm given the "observed information" and using the current estimate Λ for the parameters:

$$\begin{aligned} Q(\tilde{\Lambda}, \Lambda) &= \mathbf{E}_{\Lambda}[\log \tilde{L}(\tilde{\Lambda}) | X = m, Z = n] \\ &= \frac{\displaystyle\sum_{R \in \mathcal{R}} \log\left(e^{-|\tilde{\Lambda}|}\frac{\tilde{\Lambda}^R}{R!}\right)\frac{\Lambda^R}{R!}}{\displaystyle\sum_{R \in \mathcal{R}} \frac{\Lambda^R}{R!}}. \end{aligned}$$

Then we maximize Q over $\tilde{\Lambda}$ (for fixed Λ). It is easy to see that

$$\frac{\partial Q}{\partial \tilde{\lambda}_{kl}} = 0$$

if and only if

$$\sum_{R \in \mathcal{R}} \left(\frac{r_{kl}}{\tilde{\lambda}_{kl}} - 1 \right) \frac{\Lambda^R}{R!} = 0.$$

Hence, the only interior critical point is given by

$$\tilde{\Lambda} = f(\Lambda)$$

where, once again, f is the EM operator defined by (2.4). It is easy to see that this expression for $\tilde{\Lambda}$ gives a global maximum since Q is strictly concave in $\tilde{\Lambda}$ (for each fixed Λ). Concavity is checked by computing the Hessian:

$$\frac{\partial^2 Q}{\partial \tilde{\lambda}_{ij} \partial \tilde{\lambda}_{kl}} = \begin{cases} \dfrac{-\sum_{R \in \mathcal{R}} \dfrac{r_{kl}}{\tilde{\lambda}_{kl}^2} \dfrac{\Lambda^R}{R!}}{\sum_{R \in \mathcal{R}} \dfrac{\Lambda^R}{R!}}, & (i,j) = (k,l), \\ 0, & \text{otherwise,} \end{cases}$$

which is clearly negative definite.

References

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc.*, 39:1–38, 1977.

[2] J.J. Iannone and R.J. Vanderbei. A maximum likelihood approach to od-estimation. Technical report, Princeton University, 1994.

[3] F. Kochman, A. Murray, and D.B. West. An entropy inequality for the bi-multivariate hypergeometric distribution. *Proc. AMS*, 107:479–485, 1989.

[4] N.L. Nihan and G.A. Davis. Application of prediction-error minimization and maximum likelihood to estimate intersection o-d matrices from traffic counts. *Transportation Science*, 23:77–90, 1989.

[5] R.K.Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] Y. Vardi. Network tomography i: Estimating source-destination traffic intensities from link datas – fixed routing. Technical report, AT&T Bell Laboratories, 1994.

[7] Y. Vardi and D. Lee. From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *J.R. Statist. Soc. B*, 55:569–612, 1993.

[8] C.F.J Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11:95–103, 1983.