"A Comparison of RAS and Entropy Methods in Updating IO Tables"

S. Amer Ahmed[1] and Paul V. Preckel[2]

PRELIMINARY VERSION
PLEASE DO NOT DISTRIBUTE OR CITE WITHOUT PERMISSION
COMMENTS WELCOME

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting
Portland, Oregon, July 29-August 1, 2007

[1] S. Amer Ahmed (contact author), Ph.D. Candidate, Department of Agricultural Economics, Purdue University, 403 W. State St., West Lafayette, IN 47907. Email: saahmed@purdue.edu
[2] Paul V. Preckel, Professor, Department of Agricultural Economics, Purdue University, 403 W. State St., West Lafayette, IN 47907.

Abstract

Since the first half of the 20th century, the input-output (IO) table has been the backbone of much empirical work used to support policy analysis and develop economy-wide models. The need for accurate, up-to-date IO tables is thus essential for establishing the validity of the empirical work that follows from them. However, the construction of an IO table for any given country is an expensive and time-consuming endeavor. Current and accurate IO tables for many countries are thus often difficult to obtain on a regular basis. Once an initial IO table has been constructed, a common workaround is to collect partial information for subsequent periods, such as final demands for commodities within the economy, and then employ a Bayesian parameter estimation technique to determine values for a new IO matrix using the previous period IO table as a prior. Two such techniques to achieve this are RAS and Minimum Cross Entropy (CE).

The literature has largely ignored the question of the relative merits of these two methods. This paper uses the actual IO tables for South Korea from two distinct time periods to compare the accuracy of the RAS and CE methods. The 1995 IO table for Korea is updated to 2000 using column and row totals from the true 2000 IO table using both RAS and CE methods. The estimated IO tables are then compared to the actual 2000 IO table in order to make some observations on the relative accuracy of the methods.

The sums of squared deviations of the estimates tables from the true tables are used as the main instrument to measure deviations of the updated matrices from the true year 2000 IO matrix. It is found that the CE approach is more accurate than the RAS approach, based on the lower summed squared deviations of the elements of the CE estimated 2000 matrix from the elements of the true 2000.

The maximum absolute differences between the true and estimated tables were also calculated. It was found that the maximum absolute difference between CE-estimated table and the true posterior table was smaller than the difference between the RAS-estimated table and the true posterior.

1. Introduction

Since the first half of the 20th century, the input-output (IO) table has been the backbone of much empirical work used to support policy analysis and develop economy-wide models. The need for accurate, up-to-date IO tables is thus essential for establishing the validity of the empirical work that follows from them.

However, the construction of an IO table for any given country is an expensive and time-consuming endeavor. Current and accurate IO tables for many countries are thus often difficult to obtain on a regular basis. A common workaround is to collect partial information for subsequent periods, such as final demands for commodities within the economy, and then employ a parameter estimation technique to determine values for a new IO matrix using the previous period IO table as a prior. Two such techniques to achieve this are RAS and Minimum Cross Entropy (CE).

This paper will examine the literature associated with these two parameter estimation techniques within the context of updating IO tables using partial information. We will address the theoretical underpinnings of each method and then provide an application of each based on real world data. A comparison of an actual IO table from a subsequent time period to a table "updated" from a benchmark dataset would reveal the efficacy of these parameter estimation techniques. Using the 1995 IO table for Korea and partial information from the 2000 IO table of the same country, RAS and CE will be used to obtain estimated IO tables for 2000. The two estimated tables will then be compared to the true table for that year in an attempt to provide insight as to which method was better suited to this task.[3]

[3] GAMS code for this paper available upon request.

2. The Structure of the IO Table

This section briefly describes the generalized IO table. An understanding of the structure of the IO table will be relevant for the discussion of the Korean data to be used in Section 4.

The precursor to the modern IO table was François Quesnay's *tableau économique*[4] (Pressman, 1994). This "economic table" used a matrix framework to provide information on the sale-purchase relationships between different producers – primarily in agriculture – within an economy. The key underlying assumption to this framework was that inputs used by a given industry were related to the output by a linear and fixed coefficient production function. This basically means that the input-output relationships described by the rows and columns of the table are representative of a production technique (UN Handbook, 1999).

The industries indicated by the rows are thus producing a commodity (output) that is the input in an industry indicated by the columns. The row totals thus signify total sales of industries, while the column totals are the total costs. So, IO tables can be thought of as n x n matrices, with n being the number of industries in the economy. They are often represented as $2n$ x n, with there being twice the number of industries providing inputs. The first n rows are often for domestic inputs, while the second n are for imports. In addition to the $2n$ rows and n columns, there are additional columns (to indicate final demand) and rows (to indicate value added and taxes). The column totals together with the value added must always equal the sum of the row totals and final demand in order for the IO table to balance.

[4] Quesnay's *Tableau économique* was first published in 1758 and is now published in facsimile by the British Economic Association, London.

3. Theoretical Foundations

3.1 Updating with RAS

The RAS method is an iterative method of biproportional adjustment of rows and columns that has been independently developed by various researchers, such as Kruithoff and Sheleikhovski in the 1930s. In 1961, Stone adapted the technique for use in updating IO tables from the work of Deming and Stephan (UN Handbook, 1999).

RAS is basically an iterative scaling method whereby a non-negative matrix, M_{ij}, of dimension i x j is adjusted until its column sums and row sums equal given vectors u^* and v^* (Schneider and Zenios, 1990). This adjustment is achieved by multiplying each row by a positive constant so that the row total equals the target row total. This operation would alter the column totals. The columns would then be multiplied by constants to make their totals correspond to the target column totals. This sequence of row and column multiplication would continue until both the column and row totals converge to the target vectors.

To illustrate, let us consider the matrix M_{ij}, where M_j, is the vector of column totals. From this matrix we can obtain its matrix of coefficients[5], $A_{ij}{}^0$, as seen below.

$$A_{ij}{}^0 = M_{ij} / M_j \qquad \text{EQ (1)}$$

By pre and post multiplying this matrix by the vectors r_i and s_j, the new matrix of coefficients $A_{ij}{}^1$ is obtained. These two vectors are the vectors of target row and column totals. This matrix of coefficients is now ready to undergo a sequence of iterative multiplications, which can be seen from equations (2) through (4f). The multiplication of

[5] To avoid confusion, "matrix of coefficients" will refer to a matrix with elements divided by column totals in this paper, while "input-output table" will refer to the table of input-output value flows.

the initial coefficient matrix by the row and column multipliers gives this method its name.

$$A_{ij}^{\ 1} = r_i A_{ij}^{\ 0} s_j \qquad \text{EQ (2)}$$

The iterative process is as follows, and can be seen below. The original matrix of coefficients is multiplied by the row of target column totals, M_j^* to obtain the matrix F_{ij}.

$$F_{ij} = A_{ij}^{\ 0} M^*_{\ j} \qquad \text{EQ (3)}$$

The row totals of this matrix are represented in the vector u_i. The ratio of $u^*_{\ i}$ to u_i is the multiplier r_i. Multiplying r_i and F_{ij}, we obtain a new F_{ij}. Row vector v_j of column totals is obtained and used to calculate the multiplier s_j. F_{ij} and s_j are then multiplied. The entire sequence of operations can be seen in equations (4a-f)

$$u_i = \sum_j F_{ij} \qquad \text{EQ (4a)}$$

$$r_i = u_i^* / u_i \qquad \text{EQ (4b)}$$

$$F_{ij} = r_i F_{ij} \qquad \text{EQ (4c)}$$

$$v_j = \sum_i F_{ij} \qquad \text{EQ (4d)}$$

$$s_j = v_j^* / v_j \qquad \text{EQ (4e)}$$

$$F_{ij} = s_j F_{ij} \qquad \text{EQ (4f)}$$

The iterative process in equations (4a-f) then continues until the conditions $u_i = u^*_{\ i}$ and $v_j = v^*_{\ j}$ are met. At that point, the matrix F_{ij} is assumed to be the best estimate of the true posterior matrix M_j^*.

3.2 Updating with Minimum Cross Entropy

The CE technique – based on information theory due to Shannon (1948) – describes the entropy of a probability distribution as the measure H, where p_k is the probability of an event k occurring (Equation (5)):

$$H(p) = -\sum_{ki} p_k LNp_k \qquad \text{EQ (5)}$$

As reported by Robinson et al. (2001), from the work of Theil (1967), the entropy theory can be applied as follows. With cross entropy there is a set of *K* events with probabilities q_i of occurring. These probabilities form a prior. When confronted with a new set of information, we may desire to find a set of posterior probabilities p_i that are "close" to the prior while satisfying the restrictions embodied in the new information. Kullback and Leibler (1951) develop one measure of closeness, namely, the cross-entropy of the probability distribution described by p_i, with respect to q_i. This measure is denoted by Equation (7a):

$$-I(p:q) = -\sum_i \sum_j p_{ij} LN(p_{ij}/q_{ij}) \qquad \text{EQ (7a)}$$

Golan et al. (1994) and Golan et al. (1996) apply the Kullback-Leibler CE measure to estimate coefficients in input-output tables with the idea being to minimize *I* subject to data consistency, normalization-additivity, and new information constraints (X and y) in Equations (7b-c).

$$y_{ij} = Xp_{ij} \qquad \text{EQ (7b)}$$

$$p_{ij}'1 = 1 \qquad \text{EQ (7c)}$$

This particular paper will follow the above method as used by Robinson et al. (2001) to estimate posterior IO table coefficients. After dividing the elements of the prior IO table by its column totals, a matrix of prior "probabilities" is obtained, q_{ij}. Then, q_{ij} and p_{ij} – the matrix of posterior probabilities – are then used in the Kullback-Leibler CE measure as the objective function to be minimized (Equation (8) on the next page).

$$MIN: \sum_i \sum_j p_{ij} LN(p_{ij} / q_{ij}) \quad \text{EQ (8)}$$

The objective is minimized subject to constraints. One of them is the additivity constraint that says that all the posterior probabilities' column totals sum to one. X_j and Y_i are known column and row totals of the posterior IO tables and they form the data consistency constraints. All the constraints can be seen below in Equations (9a-c).

$$\sum_i p_{ij} = 1 \quad \text{EQ (9a)}$$

$$\sum_j p_{ij} X_j = Y_i \quad \text{EQ (9b)}$$

$$\sum_i p_{ij} X_j = X_j \quad \text{EQ (9c)}$$

Once the problem has been solved, the values of the $p_{ij} X_j$ are the values of the estimated posterior IO table.

3.3 RAS versus CE as Parameter Estimation Techniques

The literature has largely ignored the question of the relative merits of these two methods. Robinson et al. (2001) addresses this gap by comparing RAS and CE as parameter estimation techniques in the context of a 1994 SAM for Mozambique. The paper conducts simulations starting with the balanced SAM before randomly changing some row and column totals. The SAM is then updated using both the RAS and the CE methods.

From the comparison of these two methods, they found that if the focus is on column coefficients, then the CE method appears to be superior to RAS. However, if the focus is on SAM flows, then the two methods are very similar, with RAS performing slightly better. As mentioned in their paper and in McDougall (1999), the RAS and CE are equivalent measures – RAS being entropy theoretic – if the CE method uses as an

objective a single cross-entropy measure instead of attempting to use the sum of column cross-entropies.

Also, more data can be incorporated in the CE formulation, picking up changes in the flows across the matrix, and thus providing a more accurate update or estimate. RAS on the other hand is confined to using just column and row totals for the estimation technique, and would not be able to use additional information should it be available.

4. Comparison of RAS and CE with Korean IO Tables

4.1 Korea as an Application

While Robinson et al. (2001) used Monte Carlo simulations to obtain a perturbed SAM to update, we use the actual IO tables for South Korea from two distinct time periods to compare the accuracy of the RAS and CE methods. The 1995 IO table for Korea is updated to 2000 using column and row totals from the true 2000 IO table using both RAS and CE methods. The estimated IO tables are then compared to the actual 2000 IO table in order to make some observations on the relative accuracy of the methods.

The Korean IO tables for 1995 and 2000 are obtained from the Global Trade Analysis Project (GTAP). The 1995 IO table was used in the GTAP Database Version 5, while the 2000 table was used Version 6 of the database[6].

The IO tables from GTAP have the advantage of being uniform in their format and structure which makes it unnecessary to worry about cross-comparability of the two. Often IO tables used in the final GTAP Database must undergo some form of matrix balancing or "data construction" in order to provide more complete information in the final product (McDougall, 2002). These tables however were obtained before any of

[6] Documentation on these IO tables can be found at Dimaranan and McDougall (2002) and Dimaranan and McDougall (2006).

those methods were used. For more information on the format of the GTAP IO tables please refer to Huff et al. (2000) and Walmsley et al. (2002).

Two IO matrices for each year are obtained – one with taxes included, and the other without. Through addition and subtraction of column totals between the two matrices, taxes are removed from the matrix core, added together and placed in a separate row. Aside from the rows with information on value added, there are 2*n* rows for commodities, with one being for domestic goods and the other for imports. For a particular column (use *j*), the domestic and imported quantities of the commodities are added together, to collapse the 2*n* x *n* core matrix into an *n* x *n* matrix. Once that is done, the total imports for each of the *i* rows was subtracted from the row total.

At this point, the sum of all costs, value added, and taxes - the column totals – is equal to the sum of all the sales and final demand minus imports – the row totals. The IO table row and column totals equal each other.

Also, in the GTAP-ready IO tables that we use, there is a commodity known as dwellings (*dwe*) which is basically the imputed rent for residences. This faux-commodity is absent in the 1995 tables, but was included in the 2000 tables. We exclude this commodity from the 2000 table, to have a fair comparison between the tables from the two periods.

The focus of this paper is only on updating the core IO table, i.e. the fifty-six by fifty-six table of transactions between industries that is left after the final demand, value added, taxes, and negative imports rows and columns are removed. The IO core represents the intermediate demand and supply relationships between industries.

4.2 Data

The prior in this paper is the import and tax adjusted 1995 core IO table. This table, as mentioned before is fifty-six by fifty-six, for fifty-six commodities, with all values being reported in millions of Korean won. About 42.9% of the cells are zeroes. Some of these blocks of zero values can be explained by industrial structure. For example there is no oil produced in Korea, and so the entire column for oil is zeroed out. Similarly, agricultural products like food grain are not used by the construction industry. Examining the empty values in the matrix, we can see that there are three commodities that are not produced in Korea.

The 2000 core IO table – the posterior – exhibits similar features. About 40.2% of the cells are empty. Between 1995 and 2000, about 5% of cells in the IO matrix underwent a change from either a zero to a positive value, or vice versa. Some industries thus underwent technology changes such that their choices of intermediate inputs changed.

4.3 Formulation

The formulation for each method is as was described in sections 3.1 and 3.2[7]. The prior matrix is taken as given and the elements of each cell are divided by the column totals to obtain a matrix of coefficients. As posterior or target information, the column and row totals of the 2000 IO table are used.

After the estimation methods are used on the coefficient matrices to update the 1995 IO matrix to be consistent with 2000 row and column totals, the elements of each cell are multiplied by the target column totals to obtain posterior matrices that are the best estimates of the true IO table.

[7] A summary of the CE formulation is available in the Appendix.

4.4 Results

The RAS-estimated and CE-estimated matrices are compared to the true 2000 core IO matrix. As was noted before the inputs used in some industries can change dramatically between years, with some inputs dropping out of some production processes while appearing in others. So, the estimated tables may have zero elements where the true IO table has positive values. As a result, when measuring deviation between the estimated tables and the true table, it is possible that we will get some results that are very inaccurate, especially if the true table's values for the corresponding zero cells is very large.

Several different metrics are used for measurement of the methods' accuracy, the first of which is the sum of squared errors (SSE). For each estimation method, two SSE values were calculated, comparing the sum of the squared differences between the matrix of coefficients for the true and the estimated matrices of coefficients. For any given estimation method, SSE1 is the comparison of only the cells which were nonzero in both the prior and the true IO tables. SSE2 measures the deviation for all cells in the prior and true tables.

The matrices of coefficients are compared first. As will be recalled, these matrices are basically the full matrices with the cells in a column divided by the column total. By doing this, we normalize with respect to the columns to measure the relative deviation. The values can be seen in Table 1.

Table 1: Comparison of SSE Values from Matrices of Coefficients for Each Estimation Technique[8]

	SSE1	SSE2
RAS	1.017	1.113
CE	0.792	0.796

Looking down the column SSE1, it can be seen that the CE approach is superior to the RAS approach. The SSE1 measure for the CE approach is 22% smaller than the SSE1 value for the RAS approach. The results of Table 1 thus support the findings of Robinson et al. (2001) in that the CE approach provides a posterior table much closer to a true table than a posterior estimated with a RAS technique, when the matrices of coefficients are being compared.

The more general SSE2 metric considers all cells in the matrix including those that undergo a change from zero to non-zero or vice versa between the two periods. The CE measure is again superior to the RAS method.

The conclusions to be drawn from Table 1 do not change when we calculate the SSE1 and SSE2 values for not just the matrices of coefficients, but rather the true and estimated input-output tables. Table 2 compares these values and shows that the CE approach is once again superior to the RAS approach. When the input-output tables are compared, the SSE measure values for the CE approach are found to be smaller by 34%-38% than the values obtained from the RAS method.

[8] The SSE between the RAS estimated posterior and the CE estimated posterior is 0.172.

Table 2: Comparison of SSE Values from Input-Output Tables for Each Estimation Technique

	SSE1	SSE2
RAS	$3.774 \text{ X } 10^{14}$	$4.016 \text{ X } 10^{14}$
CE	$2.488 \text{ X } 10^{14}$	$2.488 \text{ X } 10^{14}$

In addition to the SSE1 and SSE2 measures, the maximum absolute differences between the true posterior and the estimated posterior tables were calculated, and can be found in Table 3. As can be seen, the CE approach yields smaller maximum absolute differences when considering both the matrices of coefficients and the estimated input-output tables.

Table 3: Comparison of Maximum Absolute Differences Between True and Estimated Posterior Tables

	Matrix of Coefficients	Input-Output Table
RAS	8088577.010	0.280
CE	5231717.586	0.275

4.5 Discussion on Specification Improvements

We might have expected the variation between the RAS-estimated and the CE-estimated tables to be smaller, since RAS is an entropy-based technique, and the two methods should provide similar results (McDougall, 1999).

One possible reason that the CE method performed better than the RAS technique is that the RAS method of biproportional adjustment is less able to deal with the cells that were zero in the prior but positive in the true posterior, and vice versa. A common strategy employed to work around this problem is to “smear” the data.

This data smear consists of taking a small percentage of the value from the non-zero elements of the row *i*, and then redistributing the collected value across the cells with zero values, maintaining row totals. This redistribution of value is equivalent to a tax on the industries using the input *i*, and as a subsidy on the industries that did not use *i* prior to the smear, but do so now.

This paper did not use any data smearing technique on the data since the purpose of this exercise was to examine the power of each estimation technique given a real world IO matrix. However, if the zero-value cells had been smeared, then it is possible that both the estimation methods might have produced posterior tables would values closer to the true IO matrix. In addition to data-smearing, there are several different ways of specifying the RAS and CE methods. This paper used a very simple iterative specification of the RAS. Most input-output analysts employ more creative specifications of the biproportional adjustment method.

5. Conclusions

RAS and CE are two parameter estimation techniques that have a long history in updating and estimating IO data. Although they share theoretical similarities, they are of varying accuracy. It is the consensus in the literature that RAS may be better than CE when flows in updated IO tables are being compared. Conversely, CE is supposed to appear more accurate than RAS when coefficient matrices of updated IO tables are considered. CE also has greater flexibility in incorporating more information in the estimation process.

This paper compares coefficient matrices of IO tables updated by both methods to the true posterior IO table, and finds that CE is more accurate than CE. The SSE values

of CE-estimated matrix were found to be significantly smaller than the SSE values of the RAS-estimated matrix. The maximum absolute difference between the true input-output table and the CE-estimated table was also found to be smaller than the maximum absolute difference between the true posterior table and the RAS-estimated table.

The accuracy of these methods are sensitive to creative strategies used by input-output analysts in dealing with these major structural shifts in the economy. Changing the cells with no value in the prior to a very small positive value while maintaining the row total in a process know as data smearing may have a significant effect on the accuracy of the techniques. Alternative specifications of the estimation methods may also provide more accurate results, and will be the focus of future work.

REFERENCES

Dimaranan, B. and R.A McDougall (eds) (2002) "Chapter 11E", *Global Trade, Assistance, and Production: The GTAP 5 Database*, Center for Global Trade Analysis, Purdue University.

Dimaranan, B. and R.A McDougall (eds) (2006) "Chapter 11D", *Global Trade, Assistance, and Production: The GTAP 6 Database*, Center for Global Trade Analysis, Purdue University.

Golan, A., G. Judge, and S. Robinson (1994) "Recovering Information from Incomplete or Partial Multisectoral Economic Data." *Review of Economics and Statistics,* 76, pp185-193.

Golan, A., G. Judge, and D. Miller (1996) *Maximum Entropy Econometrics, Robust Estimation with Limited Data*. John Wiley and Sons.

Huff, K., R.A. McDougall, and T. Walmsley (2000) "Contributing Input-Output Tables to the GTAP Data Base". GTAP Technical Paper No. 1, Release 4.2. Center for Global Trade Analysis, Purdue University.

Kulback, S and R.A. Leiber (1951) "On information and sufficiency." *Annals of Mathematical Statistics*, 4, pp.99-111.

Robinson, S¸ A. Cattaneo, and M. El-Said. (2001) "Updating and Estimating a Social Accounting Matrix Using Cross-Entropy Methods". *Economic Systems Research*, Vol 13., No. 1.

McDougall, R.A. (1999) "Entropy Theory and RAS are Friends". GTAP Working Paper No. 6. Center for Global Trade Analysis, Purdue University.

McDougall, R.A. (2002) "V5 Documentation - Chapter 19: Updating and Adjusting the Regional Input-Output Tables." Center for Global Trade Analysis, Purdue University.

Pressman, S. (1994) *Quesnay's Tableau économique : A Critique and Reassessment*. First ed., Fairfield, NJ: A.M. Kelley.

Schneider, M.H. and S.A. Zenios (1990) "A Comparative Study of Algorithms for Matrix Balancing." *Operations Research*, Vol. 38, No.3, pp. 439-455.

Shannon, C.E. (1948) "A mathematical theory of information." *Bell System Technical Journal*, 27, pp. 379-423.

Theil, H. (1967) *Economics and Information Theory* (Amsterdam, North Holland).

UN Department of Economics and Social Affairs. (1999) “Handbook of Input-Output Table Compilation and Analysis.” *Studies in Methods – Handbook of National Accounting: Series F*. United Nations, NY

Walmsley, T. and R. McDougall (2002) “V5 Documentation - Chapter 11.A: Overview of the Regional Input-Output Tables.” Center for Global Trade Analysis, Purdue University.

APPENDIX

CE Formulation

p_{ij} – matrix of coefficients of the posterior to be estimated

q_{ij} – matrix of coefficients of the prior

X_j – column totals of true posterior

Y_i – row totals of true posterior

i,j – commodities

$\sum_i \sum_j p_{ij} LN(p_{ij} / q_{ij})$ – the Kullback-Leiber entropy measure to be minimized

$\sum_i p_{ij} = 1$ – sum of column coefficients equal one.

$\sum_j p_{ij} X_j = Y_i$ – row totals of posterior IO table equal row totals of true table.

$\sum_i p_{ij} X_j = X_j$ – column totals of posterior IO table equal column totals of true table.