SAMEST: An Information-Theoretic Approach to Estimating National Accounts

Sherman Robinson and Scott McDonald

DRAFT: June 2023

Abstract

There is an extensive literature on estimating and "updating" input-output tables and Social Accounting Matrices (SAMs), including articles in recent issues of Economic Systems Research. In a survey of SAM estimation methods, Erik Thorbecke issued a challenge that it was time to develop a general stochastic approach to such estimation. Recent work using information theory and "maximum entropy" econometrics has laid the theoretical foundations for such an approach, and recent advances in computing power and solvers have made it feasible to implement methods of SAM estimation based on a general stochastic specification of measurement error in all components of the SAM accounts. The approach is Bayesian in spirit and has close links to estimation using empirical likelihood methods. The philosophy is to estimate simultaneously all cells of the SAM, which is equivalent to estimating the disaggregated national accounts, using all available information, including prior estimates of the degree of measurement error for every cell and information about any aggregates. The information may be in the form of inequalities, such as requiring cell values, aggregates, or coefficients to be positive or to be within known bounds. The recent advances in solvers and software make it feasible to implement these methods in national statistical offices.

This document is, inevitably, not a completely comprehensive or perfect description of the SAMEST programme. Ultimately the only comprehensive and perfect description of the SAMEST programme is the GAMS code.

We welcome corrections to the code and documentation and suggestions for possible improvements.

Table of Contents

1. Introduction

> "The issue of whether the SAM is deterministic or stochastic is crucial as the SAM provides the underlying data set upon which simple SAM-multiplier analyses and more complex Computable General Equilibrium Models (CGEs) are calibrated. Increasingly, these models are used to explore and simulate the impact of policies and exogenous shocks on the whole socio-economic system. An erroneous or inaccurate SAM invalidates the results obtained from these models" (Thorbecke, 2003, p 186)

At the heart of all quantitative analyses of economic systems, be it a modern macroeconomic model and/or some other form of whole economy model, will be found estimates of national accounts. Indeed, so central are such national accounts to the work of economists it is easy to forget how short the history is of (formal/institutionalised) national accounting, especially since the 'wealth of a nation' appears to be a concept that has lain at the root of economic analyses for more than two centuries (Stone, 1977, provides a brief historical review). Despite the importance of national accounts it is surprising to find how ill-informed many economists are about the issues and problems faced by national account statisticians; with the gap between economists and statisticians seeming to grow with the increasing sophistication of modern economics.[1] This is arguably a source of substantial concern since it suggests that economists are forgetting that the development of national accounts was inspired directly by developments in macroeconomics, especially the Keynesian revolution, and with it the attendant need to quantify how economic systems operate. This is not just of historical interest. In the development of national accounts there was an history of dialogue between the compilers and the users of national accounts; this dialogue had important consequences in that it has ensured that conventions for the compiling of national accounts have incorporated considerations about the use of national accounts in economic analyses. Indeed, this is one of the enduring legacies of Richard Stone's contribution to economics. This has meant that national accounts, if compiled in line with SNA guidelines, adopt definitions and conventions that ensure they can be used meaningfully as a basis for economic analyses and not solely as a mechanical accounting exercise that describes an economy at a point in time. While this dual

[1] This is not a new fresh observation - "Theorizing requires inspiration and technical know-how, while data gathering - particularly for practical implementation of large models - needs much sweat and tears, and always a large amount of time and money. No wonder we face over-production of models and underinvestment - both intellectual and financial - into compilation of the databases needed to implement them." (Leontief, 1989, p 287).

role of national accounts is among their major virtues it is all too easy for economists to forget how closely the development of national accounts was geared to the needs of economic models while simultaneously failing to recognise the difficulties confronted by national account statisticians. This compounded by a lack of recognition that the task of compiling national accounts is an exercise in estimation as opposed to measurement in the sense of a physical science. To some extent this latter failure has been compounded by the plethora of techniques that have been developed in recent years to assist in the 'balancing'/'updating'/'estimating'[1] of disaggregated national accounts. Since the most general form of disaggregated national accounts is a Social Accounting Matrix (SAM) the discussions below will be carried out in the context of a SAM although the arguments can be readily applied to many other forms of disaggregated and matrix presentations of national accounts.

It is important to recognise that the process of constructing a SAM requires the reconciliation of data that are incomplete and subject to both sampling and measurement errors. The early compilers of SAMs adopted strategies that involved confronting data from different sources with each other; taking a (subjective) view on the reliability of the different sources and then attempting to satisfy the accounting constraints of a SAM (see Pyatt *et al.*, 1977). Stone (p xxi, 1977) responded to this laudable but "laborious method" by asking whether "still better results could not be obtained by applying a formal, mathematical treatment rather than *ad hoc* manipulations to our subjective assessment of reliability" (p xxi). There appears to have been a response to this 'call' in that there are now many seemingly different techniques available for 'balancing' SAMs. More recently Thorbecke (2003) has argued that while completed SAMs are deterministic, in the sense that each cell has a unique value, it is important to recognise that the process of constructing a SAM still involves the reconciliation of data that are subject to both sampling and measurement error. In essence Thorbecke is moving the debate about strategy on beyond the mere development of mathematical techniques by arguing that it is not enough for the techniques to provide a 'mathematical' solution, but rather they must also incorporate recognition that each cell in a SAM is "an estimate arrived at on the basis of data containing sampling and measurement errors" (Thorbecke, 2003, p 185), i.e., they must also provide a statistical solution.

[1] The interpretations of these three seemingly interchangeable terms are considered further below.

The method for estimating a SAM outlined in this document is a contribution to Erik Thorbecke's challenge. It is argued that while RAS based methods, and other similar mathematical methods, achieve the objective of 'balancing' a SAM they satisfy neither Stone's nor Thorbecke's challenges. Further it is argued that while the Stone-Byron method (see Stone (1977); Byron (1978) and, for a recent summary, Round (2003)) is a major advance on RAS based methods, in that subjective judgements enter "at a second-order rather than first-order level" (Round, p 177, 2003), it also fails to fully satisfy Stone's stated objective, because, as Stone recognised at the time (p xxii), the required variance matrix "can only be based on subjective impressions of the investigator".

The method detailed below is based on information theory as applied to contexts in which the statistical problem is ill-posed since there are typically more data points to estimate than available data, i.e., there may be negative degrees of freedom. There are potentially an infinite number of solutions to a SAM estimation problem that satisfy the basic condition that expenditures by agents equal the incomes to agents, that will be manifest as an equality of row and column totals (see Gunluk-Senesen and Bates, 1987). The problem is therefore how to select from among these solutions the most likely or least unlikely solution. It is the criteria for selection that need to be guided by theory.

A legacy of the early use of the RAS method is a presumption that the objective of mathematical techniques is the generation of a matrix of transaction values consistent with a set of known row and column totals (see Bacharach *et al*., 1964). However, it is important to recognise that the limitations of such a presumption have been long recognised (see Allen and Lecomber, 1975): for instance, it is difficult to justify an assumption that the account totals are known with certainty but that the transaction values are uncertain.

The approach documented here presumes that ALL transactions in a SAM are initially estimated with error, including account totals, macroeconomic and other aggregates, and other datapoints. The estimation problem is then transformed into a process of estimating the most likely/least unlikely set of transactions that satisfy the conditions that the values of expenditures and incomes by agents are equal. This is therefore a matter of estimation NOT the mechanical balancing/updating of a SAM.

It is argued that all mathematical techniques should be evaluated on both a theoretical and empirical basis. The cross-entropy technique reported here arguably satisfies both these

criteria, but that does not mean it is the best or only technique for matrix estimation that satisfies the criteria. Since it is never possible to know that an estimated SAM is correct, it is arguable that the "updating" literature is increasingly sterile, focusing on mathematics and the uninteresting properties of various "distance" measures rather than on the question of how to use all the information available to estimate a SAM. In the "information theory" Bayesian approach, the Cross-entropy criterion is justified on axiomatic grounds: the cross-entropy metric is uniquely determined from the axioms of information theory.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of Social Accounting Matrices (SAM). This is followed in section 3 by a discussion of the SAM estimation problem by reference to the accounting and economic model considerations. In practice these problems need to be considered simultaneously and ideally the solutions should be interdependent. Section 4 details an information theory approach and reported the equations used in the SAMEST programme; this is complemented by an explanation of the operation of the error terms and weights used in the programme in section 5. The more theoretical detail in this document is supported by a User Guide.

2. Social Accounting Matrices (SAM)

The guiding principles of a SAM are the concept of the circular flow and the requirements of double entry bookkeeping. The concept of a circular flow represents a particular vision of economic systems, whereby institutions (a term that encompasses households, non-profit organisations, government, and investment) act as sellers of factor services in factor markets where producers act as purchasers, while in product markets producers act as sellers and institutions are purchasers for final consumption and other producers for intermediate products. This highlights an important distinction between Supply and Use (SUT) and Input-Output (IOT) tables and a SAM. A SAM captures the full circular flow whereas SUT and IOT only capture part of the circular flow.

A SAM is a square matrix in which each account has both a row and a column. The expenditures/payments/out-goings for each account are recorded as column entries while the incomes/receipts/in-comings for each account are recorded as row entries. As such a SAM is a single-entry form of double-entry bookkeeping. Accordingly, therefore the total expenditures by each account must be exactly equal to the total receipts for each account: hence the respective row and column sums for a SAM must equate. If a SAM is 'complete', in the sense that 'all' economic transactions are recorded, and 'consistent', in the sense that total incomes and expenditures by all agents equate, a SAM records the circular flow of an economy at a point in time. Moreover, it provides that information in an efficient and, ultimately, simple way, and in a manner that is consistent with the aggregate/macro accounts for the system. Thus, in the context of an entire economy, a SAM will contain not only the information provided by aggregate national accounts but also further details on the transactions between various groups of agents within the system. Table 2.1 is a representation a SAM which (broadly) conforms to the System of National Accounts.[1]

SAMs are generally constructed with 6 types of account and each type may contain numerous accounts: commodity/product accounts, activity/industry, factor accounts, institutional accounts (households, corporations, non-profit institutions, and government – including taxes), capital accounts (savings and investments), and rest of the world accounts.

[1] The decision to use a structure that is more reflective of the 1968 SNA is not a comment on later revisions of the SNA. It is trivial to demonstrate the structure used in Table 2.1 contains the same information as found in a SAM that adopted a structure that directly mirrored the 2008 revision of the SNA.

Table 2.1 A Stylised Social Accounting Matrix

	Commodities	**Activities**	**Factors**	**Households**	**Enterprises**	**Government**	**Capital**	**RoW**	**Total**
Commodities	0	Intermediate Inputs	0	Household Demand	??	Government Demand	Investment Demand	Commodity Exports	Total Commodity Demand
Activities	Domestic Production	0	0	0	0	0	0	0	Gross Output
Factors	0	Factor Demand	0	0	0	0	0	Factor Income from RoW	Factor Income
Households	0	0	Distributed Factor Income	Inter-Household Transfers	Distributed Dividends	Transfers	0	Remittances	Total Household Income
Enterprises	0	0	(Un)Distributed Factor Income	??	??	Fixed (Real) Transfers	0	Transfers	Total Enterprise Income
Government	Tariff Revenue VAT Other Taxes on Commodities	Indirect Taxes on Activities	Distributed Factor Income	Direct Taxes on Household Income	Direct Taxes on Enterprise Income	??	0	Transfers	Total Government Income
Capital	0	??	Depreciation	Household Savings	Enterprise Savings	Government Savings (Internal Balance)	0	Current Account 'Deficit'	Total Savings
Rest of World	Commodity Imports	0	Distributed Factor Income	Remittances	Transfers	Transfers	0	0	Total 'Expenditure' Abroad
Total	Total Commodity Supply	Total Activity Inputs	Total Factor Expenditure	Total Household Expenditure	Total Enterprise Expenditure	Total Government Expenditure	Total Investment	Total 'Income' from Abroad	

Complete and Consistent Conditions

A SAM is 'consistent' if all transactions are fully reconciled, i.e., the transaction expenditures by each agents are matched with income transactions by partner agents, and the total incomes and expenditures by all agents equate. Testing for consistency is simple: a SAM is consistent if the row and column totals equate. But a SAM can be consistent and incomplete.

A SAM is incomplete if transactions are omitted. If a SAM is incomplete, it can only be rendered consistent if other transactions have been distorted, which raises concerns about the validity of results produced using that SAM.

There is no simple way of determining if a SAM is complete. Economic theory provides some guidance in that it can be used to determine whether transactions between institutions are possible or not; the cells in the stylized SAM in Figure 2.1 that contain descriptions would be expected to contain transactions, while those with a zero ('0') entry will not contain transactions. But those labelled with question marks ('??') may or may not contain transactions. An understanding of the agents in the system can provide some insight, e.g., if the government account separately identifies central and local government then inter government transactions are likely, and if NPISH are included within the 'enterprise' accounts then final demands by 'enterprises' may be non-zero. Similarly, an understanding of an economic system may assist, e.g., remittance incomes may be known to be important to an economy and therefore some estimate must be included even if no data can be found.

Ultimately, any determination of whether a SAM is complete will depend on the knowledge of the economy provided by the compiler.

3. The SAM Estimation Problem

The estimation of a SAM requires the identification of an efficient way to incorporate and reconcile information from many different sources that may or may not have been originally collected for purposes of compiling disaggregated national accounts. In essence the cells of a SAM are unknown parameters whose values must be estimated from observed data; hence the process of compiling a SAM can be classified as a probabilistic estimation problem. But this is generally an ill-posed estimation problem since there are typically more cells/parameters to estimate than available data[1], which means there are typically negative degrees of freedom and consequently conventional statistical/econometric methods are not strictly appropriate[2].

Information theory provides one means of addressing the problem of parameter estimation as opposed to prediction. The consequent estimation principles can be defined as:

1. use all the information available; and
2. do not use, or make assumptions about, information that are not available.

In this approach it is not appropriate to make assumptions about either the error generating process or error distribution, e.g., the variance matrix of the Stone-Byron method, without evidence. Moreover, information theory provides a theoretical framework within which parameters can be estimated when data are scarce and/or incomplete.[3] This accords with Zellner's 'efficient information processing rule' and has close links with Bayesian estimation.

This exemplifies why the terms 'updating' and 'balancing' are arguably inappropriate in the context of mechanical methods used to generate new SAMs. All too often updating has

1 It may be possible to get data for certain transactions from different sources, e.g., from surveys of the agent whose income is being recorded and from the agent whose expenditure is being recorded. But it is then necessary to make a (subjective) judgement about the reliability of the different data sources; this is the approach used by early compilers, e.g., Pyatt *et. al.*, (1977) and does not obviate the problem of degrees of freedom.

2 The Stone-Byron method renders the problem susceptible to conventional econometric methods – generalised least squares – by the imposition of (subjective?) variances; hence this method side steps the problem of degrees of freedom.

3 It is tempting to argue that the issues of scarce and incomplete information are particularly problematic for developing countries where data are scarce and unreliable – measured with a lot of error. This certainly appears to be true in relation to activities within the production boundary but outside of formal market transactions and where the values of transactions must be imputed, e.g., home production for home consumption in semi-subsistence economies, but this requires a presumption that data for developed economies are necessarily more complete, which may not be the case, e.g., the withholding/suppression of data for highly concentrated industries/activities, which is often the case for the food, especially sugar, industries in the EU. Moreover, this ignores the issues of inconsistent data, which may be more important than incomplete data, and there appears to be no absolute reason to believe this issue is related to the stage of development.

referred to the derivation of a SAM for a later period primarily based upon new estimates of selected control totals, including the total incomes/expenditures for accounts, and previous transactions data[1]. On the other hand, balancing has typically referred to a removal of inconsistencies due to differences in exogenous control totals[2]. All known mechanical methods involve the use of variants of updating and balancing approaches in that they require the imposition of exogenous assumptions to render the problem solvable, e.g., the biproportionality assumption that underpins ALL variants of the RAS method. But while these methods can, and do, ensure consistency, the issue of completeness is suppressed.

Ultimately the SAM estimation problem can be regarded as constituting two related sub problems; the accounting problem, i.e., how to deal with the accounting issues, and the economic problem, i.e., how to ensure that the solution(s) to the accounting problem does not undermine the economic content of the system.

The Accounting Problem

The essence of the accounting problem is how to reconcile data from different sources. Unless the entire data gathering process for disaggregated national accounts is integrated this problem cannot be avoided since it will be necessary to use data gathered for different purposes. At its simplest the reconciliation process would involve deriving concordances between data collected using different classification schemes, e.g., trade transactions classified using Harmonised System (HS) commodity codes, production data where commodities are classified using a Standard Industrial Classification (SIC) system and household expenditure data where commodities are classified to reflect consumption patterns. These difficulties could be partially resolved by ensuring that the different surveys used a common commodity classification scheme for which requisite concordances were defined as part of the data gathering process. But this is unlikely to address all the problems, since

[1] A classic example of the updating approach was the *ex post* estimation of an IOT for 1968 using an IOT for 1963, known row and column totals for 1968 and the RAS method. A study by Lynch (1979) demonstrated the fragility of a method that used limited current information and strong assumptions about information. The improved performance of 'modified/generalised' RAS ('modified' RAS typically involves fixing some entries in the matrix and then using RAS to derive the remaining entries, see Allen and Lecomber, 1975) is an illustration of the benefits of additional information.

[2] When 'balancing' refers to the use of mechanical methods to remove minute/very small residual errors then there is arguably little reason to not use RAS or 'modified' RAS, e.g., as in Pyatt et al., (1977) where the entries in the SAM were largely reconciled by confronting data from different sources. However, when 'balancing' refers to the removal of large residual errors then mechanical methods may be seriously questionable because of the strong assumptions about the information content of different cells, especially row and column totals, that are imposed by many of these methods.

classification schemes often need to satisfy different criteria that may not always coincide, e.g., HS codes need to meet internationally defined criteria, as do ISIC codes, whereas an SIC needs to reflect the structures of a national economy.

Even if the problems presented by differences in classification schemes can be resolved this does not solve the reconciliation problem. The sources of data are typically censuses and surveys, and such data raise a series of related difficulties. Surveys face problems associated with the definition of the sample frame, which means that they may not always be perfect representations of populations, while censuses may not be complete. In addition, both surveys and censuses may fail to fully record transactions, e.g., consumers typically understate expenditures on tobacco and alcohol. But each transaction is simultaneously an expenditure by one agent and an income to another agent, hence it may be the case that there are substantive differences in the recorded values of transactions by sellers and purchasers, e.g., beverage and tobacco activities will typically record higher sales values than expenditures reported by consumers.

Consequently, it is inevitable that there will be errors in measurement and a fundamental aspect of reconciliation is to address the problem of measurement error, which is not simply a problem of mathematics.

This highlights an important point about the method reported in this paper. The necessity to confront data from different sources and make judgements about their reliability is not avoided; all the available information needs to be challenged and no information should be regarded as sacrosanct. This task may be labourious, but it remains essential. Even the development of the most sophisticated estimation techniques does not alter the requirement on data gatherers to critically evaluate the reliability of conflicting data and to consider how different data sources should be used in the process of compiling the prior estimate of the SAM. Inevitably the judgements entered into in this process risk being subjective[1], but whereas the pioneers were often required to make firm decisions about the value of the transaction/cell and the Stone-Byron method required the determination of variance and the initial value, this method only requires the determination of an initial estimate for the transaction although any additional information can and should be used.

[1] It is important to ensure that all 'subjective' judgements are confronted with evidence.

A critical consideration is the definition of information; in particular do any macroeconomic totals that may be available constitute information when compiling disaggregated national accounts. It is arguable that a theoretical ideal is that estimates of macroeconomic totals should be derived from micro level data, e.g., estimates of private consumption should be based on survey evidence and population estimates. Such an approach is arguably consistent with the principles of the SNA and the concept of using Supply and Use Tables (SUT) to 'benchmark' national accounts.[1] In the context of a national statistical agency this approach appears eminently sensible. But for non-government compilers of SAMs, it may not be practical to follow this theoretical ideal, rather such compilers may need to adopt a more pragmatic approach. Since the databases compiled by non-government agencies will rarely if ever influence the published estimates of aggregated national accounts, it is often appropriate to treat the main macroeconomic totals as binding constraints such that the disaggregated accounts are consistent with the published national accounts.[2] At first sight the requirement of consistency with exogenous macroeconomic totals may seem to make the process easier, but whereas the bottom up approach places its emphasis upon reconciling micro level data, with macroeconomic totals then being defined as deterministic aggregates, this top down approach requires that the micro level data and the macroeconomic totals must be reconciled, i.e., an additional set of constraints must be satisfied.

One advantage of this approach is that it can use as many or as few (macroeconomic) aggregates as are available, or the compiler wishes to use. This facilitates the use of the estimation method by a wide range of agencies that may or may not be acting with full access to base data.

The Economic Model Problem

The economic model problem depends on the model approach adopted; since CGE models with flexible price systems are a very general form of the class of models that use SAMs the comments here are be based on CGE models.

[1] The problems of incomplete, inaccurate, and contradictory evidence are also experienced by national account agencies when compiling national accounts, hence the estimation arguments extend to national accounts.

[2] If the disaggregated data are not consistent with the published macroeconomic data, it is easy to argue that analyses using the data are rendered invalid. Hence it may be a 'political' imperative if the database is to influence the policy making processes.

Stone's vision underpinning the System of National Accounts (SNA) always included the intention that the data systems would provide empirical content that supported mathematical models of economic systems. Central to this are three critical concepts:

1. the circular flow of transactions in an economy;
2. a coherent system of prices; and
3. the role of costs in the determination of prices.

From the circular flow stems the requirement that a SAM needs to be complete and consistent.[1] The techniques for reconciling data in a SAM can only achieve consistency, but if the SAM is incomplete but consistent then some transactions in the SAM must be distorted. Such distortions will impact on the coherence of the price system and distort the links between costs and prices. The extent to which these distortions are important will always be indeterminate.[2]

A key property of a SAM is the 'Law of One Price' (LOOP). This dictates that one column of a SAM can provide the cost information for one, and only one, price. If the column coefficients for two or more, accounts are identical then a reasonable conclusion is that the outputs of these accounts are identical (homogeneous). This feature can be obscured by attempts to be parsimonious with the dimensions of a SAM that collapse accounts, e.g., some agents may pay different (purchaser) prices for the same commodity due to differences in agent specific taxes, e.g., VAT levied only on household final demand.[3] This highlights the importance of the column coefficients for economic models, especially for price driven models.

A coherent system of prices requires that the 'rules' governing the accounting definitions for prices are transparent and recorded. The SNA price system, of basic, producer and purchaser prices, is one such system. However even within the SNA price system there are variants that need to be recognised. For instance, in standard input-output tables (IOT),

---

1 Any suggestion that a SAM is not square directly contradicts the logic of the circular flow, i.e., it requires one or more agent to have expenditures without incomes or incomes without expenditures. All known assertions that a SAM is not square derive from misrepresentations, e.g., including negative and positive entries in a row/column that sum to zero, defining aggregates that seemingly remove selected rows or columns, etc.

2 The practice of distorting a known SAM and the using a mathematical technique to estimate a SAM that can be compared to the known SAM, can only provide information about how the mathematical techniques performed in that instance.

3 If two agents can purchase the same commodity at different prices there is the possibility for profitable arbitrage. This requires that the economic system has some means to circumvent such arbitrage.

trade and transport margins are assumed to be paid by the activities, where is Supply and Use tables (SUT) they are paid by the purchasers.

A critical component of a coherent system of prices is the role of taxes as 'wedges' between the prices paid and received by agents. Since taxes are key economic policy instruments, it is critical that estimates of the tax revenues associated with individual accounts are accurate when developing a SAM. The issue of correctly assigning tax instruments in terms of the agents responsible for paying the tax, e.g., VAT is a tax on commodities not a tax on value added paid by activities, and the basis for the tax, e.g., ad valorem or specific/quantity, need to be understood if economic models are to represent correctly the tax system.

Assessing the information content of a SAM is best done in terms of the column and row coefficients, with the column coefficients taking priority. As a rule, it is not the magnitude of transactions that is important to price formation in economic models but the shares of costs, i.e., the column coefficients.[1] The magnitudes of transactions identify the relative importance of different agents. Hence when reviewing and evaluating a SAM it is good practice to check that the cost (column) and income (row) structures are coherent.

[1] Dividing all transactions in a SAM by any single number makes no change to the information content of a SAM from the perspective of a CGE model.

4. Information Theory Approach

This estimation method gets away from treating column coefficients as "analogous" to probabilities, i.e., non-negative numbers that sum to one, and explicitly specifies all error terms in stochastic terms, with a support set, probability weights, and a prior. A contribution of Golan, Judge, and Miller (1996) is to recast estimation problems into problems of estimating probabilities. Hence, the approach is a general stochastic specification, as requested by Thorbecke.

The equations for the estimation of a SAM allow the derivation of estimates of individual cells/transactions in value and coefficient form, with additive or multiplicative error terms. The formulation allows for the inclusion of multiple control totals, e.g., row and column totals, submatrix totals and national account aggregates: all control totals are defined as subject to errors that are additive or multiplicative. If the error for any transaction or aggregate is defined as ZERO the control total is binding. This flexibility with respect to control totals has costs: it adds equations and complexity to the programme.

The most complex, and least intuitive, components of the code are those that define the error terms (11 blocks of equations) and the objective function. The objective function depends on four variables that are the weights on the error terms; these weights are the probabilities defined by the error term equations. Section 5 provides details about the error specification process.

The equations are listed as they appear in the SAMEST code.

Controlling Sets

The major sets that control the equation in the programme are partly defined by the user and partly defined dynamically based the sets defined by the user.

Sets for identifying the dimensions of the SAM and a macro-SAM

sac	SAM accounts
ss	macro-SAM accounts

These sets identify the dimensions of the SAM and macro-SAM. The SAM accounts (*sac*) define the domain for most of the sets used in the estimating equations. The macro-SAM

accounts (*ss*) identify transactions from a macro-SAM that can be used as control totals when estimating the SAM.

Sets for identifying cells with specific characteristics

smcell(*sac*,*sac*)	SAM cells with abs value < *cutoff* are removed
ineg(sac,sac)	cells with negative values in data
izero(sac,sac)	cells with zero entry
nonzero(sac,sac)	cells with nonzero entry

These sets identify small prior TVs (*smcell*) and zero value TVs (*izero*) so they can be excluded from the programme. The remaining cells have non-zero (*nonzero*) or negative (*ineg*) prior TVs: these will be included in the programme although the negative prior TVs must be estimated in coefficient form, while the non-zero prior TVs can be estimated in value or coefficient form.

Sets to defined rows and columns in estimation

icol(*sac*)	columns included in estimation
icol2(*sac*)	column sums to be constrained with or without error
icolnz(*sac*)	columns with nonzero sum
irow(*sac*)	rows included in estimation
irow2(*sac*)	rows to be constrained
acol(*sac*)	additive errors for column sum constraints

These sets identify the cells to be included in the estimation. The rows (*irow*) and column (*icol*) included in the estimation are defined as those rows and columns of the prior SAM that have any non-zero TVs.

Sets for identifying cells to be estimated or fixed

icoeff(*sac*,*sac*)	cell COEFFICIENTS to be estimated
ivalue(*sac*,*sac*)	cell VALUES to be estimated
estimate(*sac*,*sac*)	cells to be estimated and not fixed
ifixv(*sac*,*sac*)	cell value fixed with no error

These sets identify the cells to be estimated in coefficient form (*icoeff*) or value form (*ivalue*). The set *ivalue* is defined as those prior TVs that are negative (*ineg*) and not zero, i.e., not *izero*, and those with non-zero column total (*icolnz*). The cells to be estimated (*estimate*) are defined as those cells with non-zero row and columns TVs (*irow* and *icol*) that are to be

estimated in value (*ivalue*) or coefficient (*icoeff*) form. If the cells are defined as members of *irow* and *icol* but not to be estimated, they are fixed (*ifix*).

Sets to define 'nature' of errors by cells

acell(*sac*,*sac*)	additive error for cells
lcell(*sac*,*sac*)	logarithmic error for cells

These sets identify whether cells to be estimated using additive (*acell*) or multiplicative/logarithmic (*lcell*) errors. All cells that are to be estimated (*estimate*) and are negative (*ineg*) are assigned to *acell*, with all other cells in estimate assigned as having logarithmic errors (*lcell*).

Prior Transaction Estimates

The programme uses several prior estimates of transaction matrices. The main matrix is the prior SAM (*P_SAM*). This matrix needs to be complete, in the sense that all cells for which there are transactions within the economy must have a non-zero value; if not then the resultant SAM will be distorted in the process of generating a SAM using any mathematical method, e.g., RAS, cross-entropy, etc. The derivation of the prior SAM is addressed elsewhere.

The second matrix is a macro-SAM. Ideally this will have been derived from national accounts data, probably first as a detailed National Accounts Matrix (NAM) and then converted to a macro-SAM format. An advantage of this route is that, if the national accounts are comprehensive, then the aggregate cells within the macro-SAM will be 'complete'. This will allow the development of the prior SAM to identify submatrices that should be nonzero.

The information in the macro-SAM can be augmented by two other databases that can provide control totals. The first is a database of aggregate totals that can include miscellaneous aggregates that involve aggregates of entries in the SAM that require data from several submatrices, e.g., GDP defined as $C + I + G + X - M$, total revenues from tax instruments. This database can be readily customised by the user, but contains a series of pre-coded definitions for (standard) aggregates. It is easily extended.

The other prior database relates to row and column totals. The programmes can derive these estimates from the prior SAM or directly from exogenous estimates provided from the Excel database. The latter option is preferred because it can encourage the user to engage more fully with the critical row and column control totals.

Estimating Equations

SAM Value and Coefficient Constraints

These equations define the values for the transaction in the SAM, in value terms either directly ($SAM_{irow,icol}$) or indirectly via the column coefficients ($COEFF_{irow,icol}$). The distinction between TVs and coefficients, and the 'nature' of errors (additive or multiplicative) are defined by the sets *acell* and *lcell*. Note how the error terms on transaction values, *ERR_TV*, are either additive or multiplicative exponential; if an error term is set to zero then the corresponding transaction is effectively fixed.

$$SAM_{irow,icol} = P\_sam0_{irow,icol} + ERR\_TV_{irow,icol} \qquad \forall ivalue \text{ AND } acell$$

$$SAM_{irow,icol} = P\_sam0_{irow,icol} * EXP\left(ERR\_TV_{irow,icol}\right) \; \forall ivalue \text{ AND } lcell$$

$$COEFF_{irow,icol} * coltarget_{icol} = coeff0_{irow,icol} + ERR\_TV_{irow,icol} * coltarget_{icol}$$
$$\forall icoeff \text{ AND } acell$$

$$COEFF_{irow,icol} * coltarget_{icol} = coeff0_{irow,icol} * EXP\left(ERR\_TV_{irow,icol}\right) * coltarget_{icol}$$
$$\forall icoeff \text{ AND } lcell$$

Column Coefficients

This equation block calculates the (column) coefficients (*COEFF*) from the transaction values (*SAM*) divided by the column total (*COLSUM*). If a *COEFF* is set it determines *SAM*; if *ivalue* sets *SAM*, then this equation determines *COEFF*.

$$COEFF_{irow,icol} * COLSUM_{icol} = \frac{SAM_{acnt,acntp}}{COLSUM_{acntp}}$$
$$\forall \textbf{ NOT } izero_{irw,icol} \text{ AND } colnz_{icol}$$

Column Sum Constraints

These equations define the values for the row (income) and column (expenditures) control totals based on the prior estimates (*rowtarget* and *coltarget*). The error terms on row and column totals, *ERR_RC*, are attached to the target control totals, i.e., it is presumed that the

control totals are measured with error.[1] NB: these equations are defined over the sets *irow2* and *icol2*, i.e., they define the row and column sums that are constrained with and without error.

$$COLSUM_{icol2} = coltarget_{icol2} + ERR\_RC_{icols,"col1"} \qquad \forall acol_{icol2}$$

$$ROWSUM_{irow2} = rowtarget_{irow2} + ERR\_RC_{irow2,"row1"}$$

Column and Row Totals

These equations calculate the row (*ROWSUM*) and column (*COLSUM*) sums from the estimated transaction matrices (*SAM*). The variables *ROWSUM* and *COLSUM* are used for the Row and Column Total Equality Constraints (see below).

$$COLSUM_{icol} = \sum_{irow} SAM_{irow,icol}$$

$$ROWSUM_{irow} = \sum_{icol} SAM_{irow,icol}$$

Row and Column Total Equality Constraints

This equation block imposes the consistency condition that total incomes (row totals) and total expenditure (column totals) must equate in the solution. Note that this condition is a constraint that must be satisfied by the matrix of bilateral transactions: it is a constraint NOT an objective.

$$COLSUM_{acbal} = ROWSUM_{acbal}$$

Macro-SAM Constraints

This block of equations is concerned with the constraints linked to a macro-SAM. The first equation calculates selected values for a macro-SAM from the estimated SAM using the mapping set *map_ss_sac* (note the index ordering is to *ss* from *sac*). The second and third equations calculate the column and row totals, respectively, for the *MACSAM*. The final,

[1] Variants of the RAS method (see Allen and Lecomber, 1975) allow the row and column controls totals to be measured with error; the use of this variant appears to be limited.

fourth, equation defines the errors for the *MACSAM* in terms of the prior macro-SAM (*macsam0*) and errors (*ERR_MSAM*).

$$MACSAM_{ssn,ssnp} = \sum_{scan,sacnp} SAM_{scan,sacnp}$$
$$\forall map\_ss\_sac_{scan,sacnp} \text{ AND } map\_ss\_sac_{ssnp,sacnp}$$

$$MACSAM_{"m\_sam\_tot",ssn} = \sum_{ss2np} MACSAM_{ss2np,ssn}$$

$$MACSAM_{ssn,"m\_sam\_tot"} = \sum_{ss2np} MACSAM_{ssn,ss2np}$$

$$MACSAM_{ss,ssp} = macsam0_{ss,ssp} + ERR\_MSAM_{ss,ssp}$$

The dimensions of the macro-SAM used in an application can be varied by the user, but does require several changes to the format of the database and the code.

Other Aggregate Constraints

The macro constraints are a means of applying aggregate control totals that are additional to those available from the macro–SAM. Typically these might be defined as aggregate across various cells in a macro-SAM, e.g., GDP equal to $C + I + G + X – M$, and total taxes on commodities, etc., which can be especially helpful in ensuring consistency with national accounts data. The aggregates, *AGGTOTAL*, are defined with error, *ERR_AGG*, as equal to an aggregate of the *SAM* where the aggregation mapping is defined by an aggregate, *aggr2*, and an aggregation (parameter) matrix, *aggagg*, of ones and zeros.

$$\sum_{irow,icol} aggagg_{irow,icol,aggr2} * SAM_{irow,icol} = AGGTOTAL_{aggr2} + ERR\_AAG_{aggr2}$$

The (Excel) database and code contain a series of pre coded aggregates, indexed *aggr* for which *aggr2* is a subset that identifies the constraints applied. The range of aggregates can be extended by adding aggregates to the Excel database.

Additional Constraint

This constraint is specific to the requirements of a SAM used to calibrate a CGE model where reexports are a problem. It defines the total exports by commodity, $\sum_w SAM_{c,w}$, as being **less**

than domestic production, $\sum_a SAM_{a,c}$, plus any export margins, $\sum_{emrgn} SAM_{emrgn,c}$, and export taxes, $\sum_{etax} SAM_{etax,c}$. The use of this constraint requires particular care when compiling the prior SAM to ensuring that reexports are excluded from the prior SAM.

$$\sum_w SAM_{c,w} = \sum_a SAM_{a,c} + \sum_{emrgn} SAM_{emrgn,c} \sum_{etax} SAM_{etax,c}$$

Error Terms

There are five discrete 'blocks' of error terms each relating to specific groups of error. The components of each error term are distinguished by a letter code that is common to the related terms.

1. '***_RC***' identifies the error terms components relating to row and column totals;
2. '***_AGG***' identifies the error terms components relating to macro aggregates;
3. '***_TV***' identifies the error terms components relating to transaction values, i.e., cells;
4. '***_MSAM***' identifies the error terms components relating to the macro-SAM;

The first block relates to the errors on row and column totals; there are two equations that define the errors, *ERR_RC*, as the weighted, *W_RC*, sum of the error support sets, *vbar_RC*, with respect to the column, *col*, and row, *row*, totals and a third that constrains the weights/probabilities to sum to one.

$$ERR\_RC_{icol2,"col"} = \sum_{jwt\_RC} W\_RC_{icol2,"col",jwt\_RC} * vbar\_RC_{icol2,"col",jwt\_RC} \quad \forall sigmay\_RC_{icol2,"col"}$$

$$ERR\_RC_{irow2,"row"} = \sum_{jwt\_RC} W\_RC_{irow2,"row",jwt\_RC} * vbar\_RC_{irow2,"row",jwt\_RC} \quad \forall sigmay\_RC_{irow2,"row"}$$

$$\sum_{jwt\_RC} W\_RC_{icol2,rwcl,jwt\_RC} = 1 \quad \forall sigmay\_RC_{icol2,rwcl}$$

The second block has two equations; the first defines the errors, *ERR_AGG*, as the weighted, *W_AGG*, sum of the error support sets, *vbar_AGG*, with respect to the various macro totals, while the second constrains the weights/probabilities to sum to one.

$$ERR\_AGG_{aggr2} = \sum_{jwt\_agg} W\_AGG_{aggr2,jwt\_agg} * vbar\_AGG_{aggr2,jwt\_agg} \quad \forall sigmay\_AGG_{aggr2}$$

$$\sum_{jwt\_agg} W\_AGG_{aggr2,jwt\_agg} = 1 \qquad \forall sigmay\_AGG_{aggr2}$$

The third block has two equations; the first defines the errors, *ERR_TV*, as the weighted, *W_TV*, sum of the error support sets, *vbar_TV*, with respect to selected transaction values, i.e., cells of the SAMs, while the second constrains the weights/probabilities to sum to one.

$$ERR\_TV_{irow,icol} = \sum_{jwt\_TV} W\_TV_{irow,icol,jwt\_TV} * vbar\_TV_{irow,icol,jwt\_TV} \quad \forall IVALUE_{irow,icol} \textbf{ OR } ICOEFF_{irow,icol} \textbf{ AND } sigmay\_TV_{irow,icol}$$

$$\sum_{jwt2} W\_TV_{irow,icol,jwt\_TV} = 1 \qquad \forall ESTIMATE_{irow,icol} \textbf{ AND } sigmay\_TV_{irow,icol}$$

The fourth block has two equations; the first defines the errors, *ERR_MSAM*, as the weighted, *W_MSAM*, sum of the error support sets, *vbar_MSAM*, with respect to (selected) aggregates of the SAM that are linked to entries in the macro-SAM, which does not need to be complete. The second constrains the weights/probabilities to sum to one.

$$ERR\_MSAM_{ss,ssp} = \sum_{jwt\_MSAM} W\_MSAM_{ss,ssp,jwt\_MSAM} * vbar\_MSAM_{ss,ssp,jwt\_MSAM} \quad \forall MACSET_{ss,ssp} \textbf{ AND } sigmay\_MSAM_{ss,ssp}$$

$$\sum_{jwt\_MSAM} W\_MSAM_{ss,ssp,jwt\_MSAM} = 1 \quad \forall MACSET_{ss,ssp} \textbf{ AND } sigmay\_MSAM_{ss,ssp}$$

Objective Function

The objective function is defined as the weighted sum of the logarithmic differences between the solution probability weights, *W_***, and the prior probability weights, '*wbar_***', where the weights are the solution probability weights, *W_***. The epsilon parameter is added to avoid zeros.

Thus, the problem of estimating a SAM is reduced to estimating the probabilities. Note how the objective function is like the objective function in RAS, which is indicative of a common heritage in information theory.

The objective function can be coded in two different, but effectively identical, ways. The first uses the CENTROPY function, which operates like a macro, and is included in the GAMS software. This is a compact representation where the inclusion of *epsilon*, as a parameter, is an option.

$$
\begin{aligned}
DENTROPY = &\left\{ \sum_{icol2,rwcl,\, jwt\_RC \forall sigmay\_RC_{icol2,rwcl}} \left[CENTROPY\left(\begin{matrix} W\_RC_{icol2,irow2,\, jwt\_RC}, \\ wbar\_RC_{icol2,irow2,\, jwt\_RC}, epsilon \end{matrix} \right) \right] \right\} \\
&+ \left\{ \sum_{aggr2,\, jwt\_AGG \forall \left(sigmay\_AGG_{aggr2} \right)} \left[CENTROPY\left(\begin{matrix} W\_AGG_{aggr2,\, jwt\_AGG}, \\ wbar\_AGG_{aggr2,\, jwt\_AGG}, epsilon \end{matrix} \right) \right] \right\} \\
&+ \left\{ \sum_{irow,icol,\, jwt\_TV \forall \left(\begin{matrix} ESTIMATE_{irow,icol} \\ \text{AND } sigmay\_TV_{irow,icol} \end{matrix} \right)} \left[CENTROPY\left(\begin{matrix} W\_TV_{irow,icol,\, jwt\_TV}, \\ wbar\_TV_{irow,icol,\, jwt\_TV}, epsilon \end{matrix} \right) \right] \right\} \\
&+ \left\{ \sum_{ss,ssp,\, jwt\_MSAM \forall \left(\begin{matrix} MACSET_{iss,ssp} \\ \text{AND } sigmay+MSAM_{ss,ssp} \end{matrix} \right)} \left[CENTROPY\left(\begin{matrix} W\_MSAM_{ss,ssp,\, jwt\_MSAM}, \\ wbar\_MSAM_{ss,ssp,\, jwt\_MSAM}, epsilon \end{matrix} \right) \right] \right\}
\end{aligned}
$$

The alternative formulation is arguably more explicit and transparent and does require the inclusion of the *epsilon* parameter. Both are included in the code with one of them commented out.

$$DENTROPY = \left\{ \sum_{icol2,rwcl,jwt\_RC \forall sigmay\_RC_{icol2,rwcl}} \left[W\_RC_{icol2,irow2,jwt\_RC} * \begin{pmatrix} LOG\left(W\_RC_{icol2,irow2,jwt\_RC} + epsilon\right) \\ -LOG\left(wbar\_RC_{icol2,irow2,jwt\_RC} + epsilon\right) \end{pmatrix} \right] \right\}$$
$$+ \left\{ \sum_{aggr2,jwt\_AGG \forall \left(sigmay\_AGG_{aggr2}\right)} \left[W\_AGG_{aggr2,jwt\_AGG} * \begin{pmatrix} LOG\left(W\_AGG_{aggr2,jwt\_AGG} + epsilon\right) \\ -LOG\left(wbar\_AGG_{aggr2,jwt\_AGG} + epsilon\right) \end{pmatrix} \right] \right\}$$
$$+ \left\{ \sum_{irow,icol,jwt\_TV \forall \left(\begin{matrix} ESTIMATE_{irow,icol} \\ \text{AND } sigmay\_TV_{irow,icol} \end{matrix} \right)} \left[W\_TV_{irow,icol,jwt\_TV} * \begin{pmatrix} LOG\left(W\_TV_{irow,icol,jwt\_TV} + epsilon\right) \\ -LOG\left(wbar\_TV_{irow,icol,jwt\_TV} + epsilon\right) \end{pmatrix} \right] \right\}$$
$$+ \left\{ \sum_{ss,ssp,jwt\_MSAM \forall \left(\begin{matrix} MACSET_{iss,ssp} \\ \text{AND } sigmay+MSAM_{ss,ssp} \end{matrix} \right)} \left[W\_MSAM_{ss,ssp,jwt\_MSAM} * \begin{pmatrix} LOG\left(W\_MSAM_{ss,ssp,jwt\_MSAM} + epsilon\right) \\ -LOG\left(wbar\_MSAM_{ss,ssp,jwt\_MSAM} + epsilon\right) \end{pmatrix} \right] \right\}$$

5. Error Specification

In general, measurement errors for various SAM statistics or coefficients are specified in the form:

$$X = \bar{X} + e \text{ or } X = \bar{X} \cdot e \tag{5.1}$$

Where X is the statistic or coefficient to be estimated, $\bar{X}$ is the prior estimate of its value, and e is the error, specified as either additive or multiplicative, depending on the views of the analyst. If the errors are multiplicative, then when using logarithms and the estimated value of the statistic or parameter can never change sign from the prior, which is often reasonable. The prior estimate of the error, e, is assumed to have a prior mean of zero for the additive case or one for the multiplicative case. Following Golan, Judge, and Miller (1996), the error e is specified as the weighted sum of a finite "error support" set, $\bar{v}$ .The weights (w) are probabilities to be estimated. The error is given by:

$$e = \sum_{jwt} w_{jwt} \bar{v}_{jwt} \tag{5.2}$$

The index jwt runs over the number of elements in the support set, the $\bar{v}_{jwt}$ are the values of the support set, and the probability weights w sum to one. This specification converts the problem of estimating errors, which are in the units of X, into a problem of estimating probabilities. The estimation procedure starts from a prior on the probability weights and revises the prior weights using all information available.

In estimation, there are two kinds of "information" that can be used. The first is information about the value of X based on information from the national accounts, including the various accounting identities and double-entry bookkeeping that are part of the defining characteristics of an economic accounting system. Second, information can be drawn from 'experts' on the nature of the errors, usually measurement error, which is incorporated in the specification of a "prior" on the error distribution.

The estimation philosophy is Bayesian in spirit: it starts from a prior distribution and then uses information to revise the prior, generating a posterior distribution. It is not necessary to specify a standard probability distribution function as a prior. It is only necessary to specify some information about the distribution through the specification of the support set and a

prior on the set of probability weights to be estimated. In contrast to standard econometric estimation procedures, which require strong assumptions about the error distribution, this approach permits incorporating widely different degrees of knowledge about the nature of the errors. The priors can range from very "uninformative" to very "informative" depending on our knowledge about the error generating process.

Seven-element uninformative prior

Following Golan, Judge, and Miller (1996), the analyst can specify an "uninformative prior" that incorporates only information about the outer bounds between which the errors must fall. In Bayesian analysis, the continuous uninformative prior is the uniform distribution between the bounds. Assuming the bounds are specified as $\pm 3s$ where s is a constant, then the prior mean is zero and the variance of the continuous uniform distribution is:

$$\sigma^2 = \frac{\left(3s - \left(-3s\right)\right)^2}{12} = 3s^2 \tag{5.3}$$

The decision is here is to specify a finite distribution with an evenly spaced, seven-element support set:

$$\bar{v}_1 = -3s \quad \bar{v}_2 = -2s \quad \bar{v}_3 = -s \quad \bar{v}_4 = 0 \quad \bar{v}_5 = +s \quad \bar{v}_6 = +2s \quad \bar{v}_7 = +3s \tag{5.4}$$

With uniform prior weights all equal to 1/7, the variance of this finite distribution is:

$$\sigma^2 = \sum_{jwt} \bar{w}_{jwt} \cdot \bar{v}_{jwt}^2 = \frac{s^2}{7}\left(9+4+1+1+4+9\right) = 4s^2 \tag{5.5}$$

The seven-element finite prior distribution provides a conservative specification of an uninformative prior. Adding more elements would more closely approximate the continuous uniform distribution. Using fewer elements yields a higher prior variance. The seven-element specification permits the estimated posterior distribution to be essentially unconstrained. In this case, with limited data, it will not be possible to recover much information about the posterior distribution.

An uninformative prior effectively reduces the number of constraints on the problem and therefore the burden on the solvers when seeking a solution. While the User Guide advocates initially using an uninformative prior, to ease finding an initial solution, the Guide

also advocates subsequently adding available information to improve the solution. Not adding available information is contrary to the imperative of using all available information.

Three-element error distribution with informative prior

In this case, assume there is more knowledge about the prior error distribution. Assume, in addition, to knowledge about the mean and upper and lower bounds, that there is prior information about the standard deviation, σ. In the case of measurement errors on aggregate data, it is assumed, in effect, that experts can provide a prior judgement on the standard error of measurement. With a general two-parameter prior distribution (mean, variance, and symmetric around zero), a finite prior can be specified with a three-element support set of $\bar{v}$s which defines the upper and lower bounds for the error distribution, and there will be three prior weights, $\bar{w}$, to be calculated. In specifying the bounds, set $S = \sigma$, so the bounds are plus and minus three times the standard deviation.

$$\begin{aligned} \bar{v}_1 &= -3\,\sigma \\ \bar{v}_2 &= \quad 0 \\ \bar{v}_3 &= +3\,\sigma \end{aligned} \tag{5.6}$$

To determine the prior weights, use the definition of the variance for a finite distribution:

$$\sigma^2 = \bar{w}_1 \cdot \left(+9\sigma^2\right) + \bar{w}_2 \cdot \left(0\right) + \bar{w}_3 \cdot \left(9\sigma^2\right) \tag{5.7}$$

Since the prior weights and support set are symmetric, $\bar{w}_{i,1} = \bar{w}_{i,3}$. Solving for the weights, $\bar{w}$, gives:

$$\begin{aligned} \bar{w}_1 &= \bar{w}_3 = \frac{1}{18} \\ \bar{w}_2 &= 1 - \bar{w}_1 - \bar{w}_3 = \frac{16}{18} \end{aligned} \tag{5.8}$$

In this case, the specification of a three-element support set permits recovery of only limited information about the posterior distribution.

This prior information can be especially useful when detailed survey data, e.g., household income and expenditure surveys and labour force surveys, are available that can be used to derive point estimates and associated standard deviations/variances. This class of

information is akin to the Stone-Byron method where point estimates and variances are assigned to all transactions, except in this method it can be applied to a subset of transactions. Note however that the least squares and cross-entropy estimation methods are different.

Five-element error distribution with informative prior

For the case of a five-parameter support set for an informative prior, there are five prior weights, $\bar{w}$, to be specified. This effectively incorporates more information about the prior error distribution—more moments, including variance, skewness, and kurtosis. Assuming in addition to a prior mean and variance, a prior normal distribution is specified; then the prior on skewness is zero and kurtosis is $3\sigma^4$. In this case, the prior weights, $\bar{w}$, are calculated so that:

$$\sum_{jwt} \bar{w}_{jwt} \cdot \bar{v}_{jwt}^4 = 3\,\sigma^4 \tag{5.9}$$

The prior weights and support sets are also symmetric, so the prior on all odd moments is zero (mean and skewness are zero). Choose $\pm$ 1.5 times the standard deviations for $\bar{v}_{i,2}$ and $\bar{v}_{i,4}$ (which is arbitrary but follows the usual practice of evenly spacing the support set). In this case:

$$\begin{aligned}
\bar{v}_1 &= -3.0\,\sigma \\
\bar{v}_2 &= -1.5\,\sigma \\
\bar{v}_3 &= \quad 0 \\
\bar{v}_4 &= +1.5\,\sigma \\
\bar{v}_5 &= +3.0\,\sigma
\end{aligned} \tag{5.10}$$

Using the definition of variance and kurtosis:

$$\begin{aligned}
\sigma^2 &= \bar{w}_1 \cdot \left(9\,\sigma^2\right) + \bar{w}_2 \cdot \left(2.25\,\sigma^2\right) + \bar{w}_3 \cdot \left(0\right) + \bar{w}_4 \cdot \left(2.25\,\sigma^2\right) + \bar{w}_5 \cdot \left(9\,\sigma^2\right) \\
3\sigma^4 &= \bar{w}_1 \cdot \left(81\,\sigma^4\right) + \bar{w}_2 \cdot \left(\frac{81}{16}\sigma^4\right) + \bar{w}_3 \cdot \left(0\right) + \bar{w}_4 \cdot \left(\frac{81}{16}\sigma^4\right) + \bar{w}_5 \cdot \left(81\,\sigma^4\right)
\end{aligned} \tag{5.11}$$

Given that $\bar{w}_1 = \bar{w}_5$ and $\bar{w}_2 = \bar{w}_4$ by symmetry, there are two equations in two unknowns:

$$\begin{aligned} 18\,\bar{w}_1 + 4.5\,\bar{w}_2 &= 1 \\ 162\,\bar{w}_1 + \frac{81}{8}\bar{w}_2 &= 3 \end{aligned} \tag{5.12}$$

Solving for the prior weights, noting that they must sum to one:

$$\bar{w}_1 = \bar{w}_5 = \frac{1}{162} \qquad \bar{w}_2 = \bar{w}_4 = \frac{16}{81} \qquad \bar{w}_3 = \frac{48}{81} \tag{5.13}$$

These parameters determine the prior distribution. The estimation procedure yields posterior estimates of the error distribution, and the five-element specification of the support set permits estimation of four moments of the posterior distribution (mean, variance, skewness, and kurtosis).

References

Allen, R.I.G. and Gossling, W.F., (eds) (1975). Estimating and Projecting Input-Output Coefficients. London: Input-Output Publishing Company.

Allen, R.I.G. and Lecomber, J.R.C., (1975). 'Some Tests on a Generalised Version of RAS' in Allen, R.I.G. and Gossling, W.F., (eds) Estimating and Projecting Input-Output Coefficients. London: Input-Output Publishing Company.

Lecomber, J.R.C., (1975). 'A Critique of Methods of Adjusting, Updating and Projecting Matrices' in Allen, R.I.G and Gossling, W.F., (eds) Estimating and Projecting Input-Output Coefficients. London: Input-Output Publishing Company.

Leontief, W.W., (1989). 'Input-Output Data Base for Analysis of Technical Change', Economic Systems Research, Vol 1, pp 287-295.

Stone, R., (1956). Quantity and Price Indexes in National Accounts. Paris: OEEC.

Stone, R., (1961). Input-Output and National Accounts. Paris: OEEC.

Stone, R., Bates, J.M. & Bacharach, M., (1963). 'Input-Output Relationships 1954-66', Vol 3 in Stone, R, (ed) A Programme for Growth. London: Chapman & Hall.

Bacharach, M., (1970). Biproportional Matrices and Input-Output Change. Cambridge: Cambridge University Press.

Byron, R.P., (1976). 'The Estimation of Large Social Account Matrices', Journal of the Royal Statistical Society, A, Vol 141, pp 359-367.

Byron, R.P., (1996). 'Diagnostic Testing and Sensitivity Analysis in the Construction of Social Accounting Matrices', Journal of the Royal Statistical Society, A, Vol 159, pp 133-148.

Dervis, K., de Melo, J. and Robinson, S., (1982). General Equilibrium Models for Development Policy. Washington: World Bank.

Golan, A., Judge, G. and Miller, D., (1996). Maximum Entropy Econometrices. Chichester: John Wiley & Sons.

Golan, A., Judge, G. and Robinson, S., (1994). 'Recovering Information from Incomplete or Partial Multisectoral Economic Data', Review of Economics and Statistics, Vol 76, pp 541-549.

King, B.B., (1985). 'What is a SAM?', in Pyatt, G. and Round, J.I. (ed.), Social Accounting Matrices: A Basis for Planning. Washington: World Bank.

Lynch, R.G., (1979). 'An Assessment of the RAS Method for Updating Input-Output Tables', in Sohn, I., (ed) Readings in Input-Output Analysis. Oxford: Oxford University Press.

Pyatt, G., (1987). 'A SAM Approach to Modelling', Journal of Policy Modeling, Vol. 10:327-352.

Robinson, S., Cattaneo, A., and El-Said, M., (1998). 'Estimating a Social Accounting Matrix Using Cross Entropy Methods', mimeo.

Stone, R., (1962). A Social Accounting Matrix for 1960. Volume 2 in A Programme for Growth. London: Chapman and Hall.

Stone, R., (1974). 'Forward' in Pyatt, G., Roe, A.R. and Associates, Social Accounting Matrices for Development Planning with Special Reference to Sri Lanka. Cambridge: Cambridge University Press.

Stone, R., (1985). 'The Dissaggregation of the Household Sector in the National Accounts', in Pyatt, G. and Round, J.I. (ed.), Social Accounting Matrices: A Basis for Planning. Washington: World Bank.

Stone, R., Bates, J.M. and Bacharach. M., (1963). Input-Output Relationships 1954-66. Volume 3 in A Programme for Growth. London: Chapman and Hall.

UN (1968). A System of National Accounts. Studies in Methods, Series F, Rev 4. New York: United Nations.

UN (1993). System of National Accounts. New York: United Nations.