Entropy **2014**, *16*, 1243-1271; doi:10.3390/e16021243

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

A Bayesian Approach to the Balancing of Statistical Economic Data

João F. D. Rodrigues

Instituto Superior Técnico, Unversidade de Lisboa, IN+, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal; E-Mail: joao.rodrigues@tecnico.ulisboa.pt; Tel.: +351-21-8417374

Received: 16 January 2014; in revised form: 12 February 2014 / Accepted: 17 February 2014 / Published: 26 February 2014

Abstract: This paper addresses the problem of balancing statistical economic data, when data structure is arbitrary and both uncertainty estimates and a ranking of data quality are available. Using a Bayesian approach, the prior configuration is described as a multivariate random vector and the balanced posterior is obtained by application of relative entropy minimization. The paper shows that conventional data balancing methods, such as generalized least squares, weighted least squares and biproportional methods are particular cases of the general method described here. As a consequence, it is possible to determine the underlying assumptions and range of application of each traditional method. In particular, the popular biproportional method is found to assume that all source data has the same relative uncertainty. Finally, this paper proposes a simple linear iterative method that generalizes the biproportional method to the data balancing problem with arbitrary data structure, uncertainty estimates and multiple data quality levels.

Keywords: statistical economic data; data balancing; Bayesian approach; relative entropy minimization; uncertainty; data quality

---

1. Introduction

In the compilation of statistical economic data, such as a census-based Input-Output (IO) table or a social-accounting matrix (SAM), it is often the case that the data is not balanced, i.e., row and column sums do not add up [1]. Furthermore, data balancing is important in practical applications such as updating or regionalizing IO tables, or decomposing proximate causes of economic change [2–4]. So as more countries develop IO tables with greater regularity and regional SAMs for computable general

equilibrium (CGE) modeling are used more, the use of balancing techniques will undoubtedly rise as well.

As Lahr and de Mesnard [5] note, many alternative formulations do exist that can perform a table balancing. Empirical work demonstrating the merits and costs of the various approaches are not always convincing. Indeed, because no theory of optimal IO data processing exists, there is no way to figure out a priori which particular technique will work best under particular circumstances.

In this paper I intend to develop a theory for balancing elements in input-output tables based on the theory of Bayesian inference of Jaynes [6]. This approach has appeal because it is based on first principles and does not rely on ad-hoc reasoning. Using it I am therefore able to prove which numerical algorithm is best suited for a given set of uncertainty parameters in a set of IO accounts.

The present paper addresses the problem of IO data balancing under the following conditions:

- The constraints are not necessarily biproportional but can take arbitrary structure.
- There is some degree of uncertainty affiliated with the values of IO elements.
- The IO elements may come from different sources with differing degrees of data quality.

In the classical biproportional or RAS problem [5,7] the intermediate inputs in a matrix are adjusted while row and column sums are fixed. When arbitrary structure is considered, every element in the data set may be constrained to be the sum of a subset of all other elements in the data set.

Data uncertainty is an estimate of the empirical error associated with each numerical datum. In contrast to biproportional balancing methods [5] and their variants [8], in this paper it is considered that every datum is characterized both by a best guess and by an uncertainty estimate. Some optimization methods [9], such as least-squares methods [10], allow the use of uncertainty information during balancing but provide no general rule to determine uncertainty when that information is initially absent.

Finally, data balancing problems frequently involve the combination of data from several sources with potentially different degrees of quality. For example, in the classical table update problem [5], there is an initial estimate from the previous year for interior points (low quality data) and row and column sums for the present year (high quality data). In practice the data update problem combines data from multiple sources and differing degrees of trustworthiness [11–13]. The present paper deals with the general problem of combining data with differing degrees of quality (e.g., data from national statistical offices, from international organizations, survey data, *etc.*).

Currently, there is no data balancing method that addresses all of these issues, even though all of them arise in the compilation of multi-regional IO models. In this paper, this problem is solved using concepts and techniques of Bayesian inference [6].

Conventional methods address the balancing problem by imposing constraints on data interpreted as real numbers. In contrast, in a Bayesian framework, data are interpreted as random variables, and constraints are imposed on their first and second moments (best guess and uncertainty). Application of relative entropy minimization leads to an analytical solution.

Unfortunately, the analytical solution is impractical, so a series of numerical approximations is derived, whose validity depends on the amount of uncertainty information initially available. After this derivation conventional data balancing methods are reviewed and a one-to-one correspondence between the conventional methods and the numerical approximations is identified.

The existence of a one-to-one correspondence between Bayesian and conventional methods means that it is possible to identify the underlying assumptions of conventional methods. In particular, the popular RAS method assumes all data to have the same relative uncertainty.

Therefore, the Bayesian linear algorithm (recommended for most practical applications) turns out to be a generalization of the classical RAS method to the situation of arbitrary structure, uncertainty information and data quality hierarchy.

The paper proceeds as follows. Section 2 derives the general solution and numerical simplifications of the Bayesian data balancing method. Section 3 reviews conventional methods and compares them to the Bayesian methods. Section 4 concludes and the Appendix A reports auxiliary material.

2. Bayesian Methods

2.1. Problem Formulation

This paper addresses the problem of balancing an IO table with arbitrary structure, uncertainty estimates and multiple data sources. These three properties are modeled as follows.

An arbitrary structure is formalized by considering that the IO data is arranged in a vector $\mathbf{t}$ of length n_T and is subject to n_K *accounting identities* of the form:

$$0 = \sum_{j=1}^{n_T} G_{ij} t_j + k_i, \tag{1}$$

where k_i is a *numerical constraint* and each G_{ij} can take values -1, 0 or 1. The accounting identities can be arranged in a constraint vector $\mathbf{k}$ and a *concordance matrix* $\mathbf{G}$, such that:

$$\mathbf{0} = \mathbf{G}\mathbf{t} + \mathbf{k}, \tag{2}$$

where $\mathbf{0}$ is a vector of zeros and $\mathbf{t}$ is the balanced *posterior*. The starting point for the balancing procedure is the unbalanced *prior*, $\boldsymbol{\theta}$ for which:

$$\mathbf{0} \neq \mathbf{G}\boldsymbol{\theta} + \mathbf{k}. \tag{3}$$

In a nutshell, the data balancing problem with arbitrary structure is as follows: initially there is knlowledge of $\mathbf{G}$, $\mathbf{k}$ and $\boldsymbol{\theta}$, satisfying Equation (3) and the goal is to determine the $\mathbf{t}$ which satisfies Equation (2) and satisfies some additional properties.

For the purpose of this paper it is considered that every entry of the data vector is positive. Appendix A.1 shows how to deal with negatives and zeros in the original IO table.

To illustrate the construction of the concordance matrix and constraint vector, consider the case of a 2×2 matrix $\mathbf{Z}$, with known row and column sums, $\mathbf{Z}\mathbf{e} = \mathbf{x}^{\mathbf{R}}$ and $\mathbf{Z}'\mathbf{e} = \mathbf{x}^{\mathbf{C}}$, where the row and column sums are $\mathbf{x}^{\mathbf{R}}$ and $\mathbf{x}^{\mathbf{C}}$, and $\mathbf{e}$ is a vector of ones. This problem is formulated with $n_T = 4$ numerical data and $n_K = 4$ accounting identities, as:

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}; \quad \mathbf{t} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ Z_{11} \\ Z_{22} \end{bmatrix}; \quad \mathbf{k} = \begin{bmatrix} x_1^R \\ x_2^R \\ x_1^C \\ x_2^C \end{bmatrix}. \tag{4}$$

The system described by Equation (4) is the conventional RAS problem.

The handling of uncertainty estimates requires the formalization of the stochastic properties of the IO data. Following Weise and Woger [14], who apply concepts of Bayesian inference [6] to the problem of measurement errors, in this paper it is considered that each IO datum is subject to empirical measurement errors and is therefore described by a random variable.

Thus, the prior $\boldsymbol{\theta}$ is characterized by a probability distribution $\pi(\mathbf{q})$, which expresses the degree of belief that the inaccurately known prior takes realization $\mathbf{q}$. The prior *best guess* or expectation vector is $\boldsymbol{\mu}$, the prior *uncertainty* or standard-deviation vector is $\boldsymbol{\sigma}$, and the prior *correlation* matrix is $\mathbf{P}$. The posterior $\mathbf{t}$ is in turn characterized by a probability distribution $p(\mathbf{q})$, the posterior best guess vector is $\mathbf{m}$, the posterior uncertainty vector is $\mathbf{s}$ and the posterior correlation matrix is $\mathbf{R}$. The prior and posterior covariance matrices are, respectively, $\boldsymbol{\Sigma} = \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}$ and $\mathbf{S} = \hat{\mathbf{s}}\mathbf{R}\hat{\mathbf{s}}$, where $\hat{}$ denotes diagonal matrix.

This paper considers that the probability distribution, best guess, uncertainty and correlations of the prior are known. The best guess, m_i, and uncertainty, s_i, of a numerical datum are referred to as *observables*, to distinguish them from the corresponding parameters of the truncated Gaussian distribution, $\tilde{m}_i$ and $\tilde{s}_i$, that will appear in Section 2.2.

Finally, the problem of combining different data sources is formalized with the concept of *data quality*. That is, this paper considers that besides quantitative uncertainty information the source data is also characterized by a qualitative ranking, $\mathbf{h}$, which indicates how trustworthy that data point is relative to others.

The ranking of data quality is used to solve the problem of conflicting constraints. Essentially, the present paper suggests that data of lower quality should be balanced while keeping data of higher quality fixed as constraints. But if a balanced solution cannot be found, then the "constraints" become adjustable.

To illustrate this concept, consider the RAS problem described in Equation (4). In this case the entries of the $\mathbf{Z}$ matrix are of lower quality than the row and column sums. Thus, the general problem, taking into account data quality can be formulated with $n_T = 8$ numerical data and $n_K = 4$ accounting identities, as:

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}; \quad \mathbf{t} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ Z_{11} \\ Z_{22} \\ x_1^R \\ x_2^R \\ x_1^C \\ x_2^C \end{bmatrix}; \quad \mathbf{h} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, \tag{5}$$

and $\mathbf{k} = \mathbf{0}$. The problem defined by Equation (4) has a single level of data quality (where interior points are adjusted while row/column sums are fixed), whereas the problem defined by Equation (5) has two data quality levels, allowing for row and column sums to be adjusted too.

For clarity of exposition the remainder of this section is as follows. The data balancing problem with just two quality levels (numerical data and numerical constraints) is studied in Section 2.2. Data quality

and the construction of numerical constraints is addressed in Section 2.3. Finally, Section 2.4 presents numerical approximations.

2.2. Analytical Solution

Bayesian inference was first developed by Laplace [15] and later expanded by others, such as Jeffreys [16] and Jaynes [6,17]. According to the Bayesian paradigm, a probability is a degree of belief about the likelihood of an event, and should reflect all relevant available information about that event. If more information about the event becomes available, then the *prior* probability must be updated to a *posterior* probability.

In the data balancing problem the goal is to update a *probability distribution*, under the guiding principle that *the best inference is the one which takes into account all available information and no other*. This principle is operationalized by searching for a posterior distribution that is as close as possible to the prior (in an information sense) and that satisfies the accounting identities, expressed in terms of moment constraints.

That is, if a discrete distribution is considered, the goal is to obtain a posterior, $p(q_j)$, when both a prior, $\pi(q_j)$, and moment constraints are known, by *minimizing relative entropy* [18,19]. The Lagrangean is:

$$L = \sum_{j=1}^{n_L} p(q_j) \log\left(\frac{p(q_j)}{\pi(q_j)}\right) + \sum_{i=1}^{n_M} \lambda_i \left(M_i - \sum_{j=1}^{n_L} (q_j)^i p(q_j)\right). \tag{6}$$

The first term on the right hand side of Equation (6) is the entropy of $p(q_j)$ relative to $\pi(q_j)$, and the second term is the set of moment constraints. n_L is the number of discrete realizations, n_M is the number of moment constraints and M_i is the i-th moment (e.g., the first moment is the best guess, the second moment is the variance). The solution of relative entropy minimization takes the form:

$$p(q_j) = \frac{\pi(q_j)}{Z} \exp\left(\sum_{i=1}^{n_L} \lambda_i (q_j)^i\right). \tag{7}$$

Z is a normalization factor to convert relative probabilities into absolute ones. According to Robinson *et al.* [20] (p. 52), the solution of relative entropy minimization "is analogous to Bayes' Theorem, whereby the posterior distribution, $p(q_i)$, is equal to the product of the prior distribution, $\pi(q_i)$, and the likelihood function (probability of drawing the data given parameters being estimated), $\exp(\sum_{i=1}^{n_L} \lambda_i (q_j)^i)$, dividing by a normalization factor, Z."

As reviewed in Section 3.1, there is a class of conventional cross-entropy methods in which an IO datum is treated as a scalar, t_i, and so the constraints take the form of Equation (1). That formalization is radically different from the Bayesian interpretation followed here, in which a numerical datum is conceptualized as a random variable. To our knowledge no data balancing method using the Bayesian interpretation of IO data has ever been proposed, although Golan *et al.* [21] offer a bridge between the two interpretations (datum as scalar and datum as random variable) through the concept of generalized cross entropy (see Section 3.1).

According to the Bayesian paradigm the best solution to the data balancing problem should take all available information into account. This information are the constraints of the first and second moments of the numerical data. Appendix A.2 shows that the constraints take the matrix form:

$$\mathbf{0} = \mathbf{G}\mathbf{m} + \bar{\mathbf{m}}; \tag{8}$$

$$\mathbf{0} = \text{diag}\left(\mathbf{G}\mathbf{S}|\mathbf{G}|'\right) + \bar{\mathbf{s}}^{\mathbf{2}}, \tag{9}$$

where $\bar{\mathbf{m}}$ and $\bar{\mathbf{s}}^{\mathbf{2}}$ are the vectors of best guess and uncertainty constraints. The construction of these vectors is explained in Section 2.3.

Appendix A.3 shows how the introduction of these constraints in the cross-entropy minimization problem leads to the solution:

$$\tilde{\mathbf{S}}^{-1} = \tilde{\mathbf{\Sigma}}^{-1} + \mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|; \tag{10}$$

$$\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{m}} = \tilde{\mathbf{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}} + \mathbf{G}'\boldsymbol{\alpha}, \tag{11}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the first and second moment Lagrange parameters. Taken together, Equations (10) and (11), Equations (8) and (9) define the analytical solution of the Bayesian data balancing method. Note however that Equations (10) and (11) contain symbols adjoined with ˜ (Gaussian parameters) while Equations (8) and (9) do not.

When relative uncertainty, σ_j/μ_j or s_j/m_j, is low, then the Gaussian parameter and the observable are identical. When relative uncertainty is high, the best guess Gaussian parameter tends to $-\infty$ and the uncertainty Gaussian parameter tends to ∞. There is no closed-form expression between observables and Gaussian parameters in the multivariate case.

2.3. Data Quality

This subsection introduces the concept of data quality, which determines the sequence in which the data balancing procedure is implemented and how numerical constraints are constructed. As described in Section 2.1, a key motivation for the present work is the possibility to incorporate information on the quality of source data directly in the balancing method.

Thus, consider that each numerical datum i is characterized by an integer-valued number h_i which indicates its *quality*. That is, if datum i is more trustworthy than datum j, then $h_i > h_j$. Section 2.1 gives the example of a 2 by 2 RAS problem, in which the row and column sums were assumed to have higher quality than interior points. (The choice of integer values for the entries of $\mathbf{h}$ is for convenience only, any ordinal ranking such as a, b, c, *etc.*, would work as well.)

Issues of data quality inevitably arise in the compilation of IO tables from multiple data sources. If a practitioner wishes to construct a table combining official data from a national statistical office with survey data and data collected by third parties, it is likely that discrepancies between the different datasets will arise. When removing those discrepancies (the purpose of data balancing), it is natural that the method should allow the practitioner to use a qualitative measure of how trustworthy the different datasets are, relative to one another.

Data quality, h_i, should not be confused with uncertainty estimate, s_i. The latter is a *quantitative* expression of how trustworthy the best guess, m_i, is. The former is a *qualitative* expression of how trustworthy (m_i, s_i) are, in relation to other source data (m_j, s_j) where $j \neq i$.

The present paper suggests to incorporate data quality in the balancing problem by considering that *higher quality data is fixed while lower quality data is being balanced*, and only if a balanced solution cannot be found is higher quality data adjusted too.

Consider that among the n_T numerical data there are Q data quality levels, and the numerical data are indexed by increasing level of data quality. That is, all points in the range $(n_{L-1}+1, n_L)$ have data quality of level L, where $n_0 = 0$ and $n_Q = n_T$. The method searches for a balanced solution of quality level L, by holding fixed all data points $j > n_L$. The method starts with $L = 1$ and moves up until a solution is found. In the worst-case scenario, a solution always exists when $L = Q$ and all data can be balanced.

That is, in the data balancing problem at level L, the vectors of numerical data and the columns of the concordance matrix are truncated from n_T to n_L, and the posterior moment constraints (Equations (8) and (9)) become:

$$\begin{aligned} \mathbf{0} &= \mathbf{G}(L)\mathbf{t}(L) + \mathbf{k}(L); \\ \mathbf{0} &= \mathbf{G}(L)\mathbf{m}(L) + \bar{\mathbf{m}}(L); \\ \mathbf{0} &= \operatorname{diag}\left(\mathbf{G}(L)\mathbf{S}(L)|\mathbf{G}(L)|'\right) + \bar{\mathbf{s}}^{\mathbf{2}}(L). \end{aligned} \tag{12}$$

The numerical constraints $\mathbf{k}(L)$, introduced in Section 2.1, are therefore an aggregation of higher quality data, for the particular balancing problem of level L. The constraint best guess, $\bar{\mathbf{m}}(L)$, and variance, $\bar{\mathbf{s}}^{\mathbf{2}}(L)$, vectors, introduced in Section 2.2, are defined as:

$$\begin{aligned} \mathbf{k}(L) &= \mathbf{G}(Q)\boldsymbol{\theta}(Q) - \mathbf{G}(L)\boldsymbol{\theta}(L); \\ \bar{\mathbf{m}}(L) &= \mathbf{G}(Q)\boldsymbol{\mu}(Q) - \mathbf{G}(L)\boldsymbol{\mu}(L); \\ \bar{\mathbf{s}}^{\mathbf{2}}(L) &= \operatorname{diag}\left(\mathbf{G}(Q)\boldsymbol{\Sigma}(Q)|\mathbf{G}(Q)|'\right) - \operatorname{diag}\left(\mathbf{G}(L)\boldsymbol{\Sigma}(L)|\mathbf{G}(L)|'\right), \end{aligned} \tag{13}$$

where $\mathbf{G}(L)$, $\boldsymbol{\theta}(L)$, $\boldsymbol{\mu}(L)$ and $\boldsymbol{\Sigma}(L)$ are, respectively, the concordance matrix, the prior random vector, the prior best guess vector and the prior covariance matrix at quality level L. It follows that at the highest quality level, Q, the numerical constraints are zero, $\mathbf{0} = \mathbf{k}(Q) = \bar{\mathbf{m}}(Q) = \bar{\mathbf{s}}^{\mathbf{2}}(Q)$.

The solution at the current quality level is incorporated in the prior of the next quality level: $\mu_j(L+1) = m_j(L)$, $\sigma_j(L+1) = s_j(L)$ and $\rho_{jk}(L+1) = r_{jk}(L)$ for $j = 1, \ldots, n_L$ and $k = 1, \ldots, n_L$.

A word of caution is necessary. If the assignment of data quality is incorrect, it is possible that the problem becomes ill posed. As a general rule, the user should always check if the results are meaningful: the Bayesian data balancing method can only provide a good solution if good data is provided. This is not a handicap but an advantage of the method, because it is warning the practitioner that the initial assignment of data quality is incorrect. This behavior is in agreement with the suggestion of Jaynes [6] that a Bayesian inference robot should apply rules uncritically, so that if an absurd outcome emerges, it is easy to identify the error in the problem formulation.

2.4. Numerical Approximations

The analytical solution of Section 2.2 requires the analytical conversion from the multivariate truncated Gaussian parameters to observables [22,23] and matrix inversions [24], operations which are far from trivial.

In this subsection a series of numerical approximations is reported, whose validity depends on how well source data uncertainty is characterized, which will in turn affect the value of correlations.

In practical applications, it happens frequently that an accounting identity (introduced in Section 2.1) contains only one entry $G_{ij} = -1$ and several entries $G_{ij} = 1$. In this paper the former is referred to as an *aggregate* datum and the latter as *disaggregate* data.

If there is a good characterization of all source data uncertainties, then the generalized least squares algorithm should be used. If there is a good characterization of disaggregate data but a poor one of aggregate data, then the weighted least squares algorithm or the linear algorithm should be considered. Finally, if there is a poor characterization of all uncertainties, the proportional algorithm should be preferred.

All of these algorithms are iterative and at each step the best guess displacement must be kept small and relative uncertainty constant.

2.4.1. The GLS Algorithm

The *generalized least-squares (GLS) algorithm* is obtained under two simplifying assumptions.

The first and strongest assumption is to replace the truncated multivariate Gaussian with the non-truncated Gaussian, while still imposing that observable uncertainty is bound by observable best guess, $0 < \sigma \leq \mu$ and $0 < s \leq m$ (Section 2.1). As shown in Section 3, the algorithms derived from this simplification turn out to be generalizations of the most used conventional data balancing methods. Thus, if in the future someone proves that the numerical algorithms proposed here are bad approximations of the analytical solution, that would imply the data balancing practice of the past 50 years is also wrong. If the present paper is the catalyst of such a revolutionary discovery, that alone is a valid contribution to the literature.

The second assumption is to consider that best guesses, $\boldsymbol{\mu}$, are known more accurately than uncertainties and correlations, $\boldsymbol{\sigma}$ and $\mathbf{P}$, and so uncertainties and correlations should be adjusted before best guesses. That way, if the best guesses are initially balanced, they will remain unchanged.

Thus, if second-order data is initially balanced, $\mathbf{S} = \boldsymbol{\Sigma}$, Equation (11) simplifies to:

$$\mathbf{m} = \boldsymbol{\mu} + \mathbf{S}\mathbf{G}'\boldsymbol{\alpha}. \tag{14}$$

The combination of Equations (8) and (14) determines the best guess Lagrange multipliers, $\boldsymbol{\alpha}$, as the solution of:

$$\left(\mathbf{G}\mathbf{S}\mathbf{G}'\right)\boldsymbol{\alpha} = -\left(\mathbf{G}\boldsymbol{\mu} + \bar{\mathbf{m}}\right). \tag{15}$$

Equations (14) and (15) represent a generalized least-squares (GLS) solution, which is valid if a balanced set of covariances has been found.

This problem may lack a solution if some accounting identities are linearly dependent and the corresponding numerical constraints are inconsistent. This case can be addressed by finding the minimum-norm solution, i.e., the $\boldsymbol{\alpha}$ which minimizes $||\boldsymbol{\alpha}||_2$ and:

$$||\left(\mathbf{G}\mathbf{S}\mathbf{G}'\right)\boldsymbol{\alpha} - \left(\mathbf{G}\boldsymbol{\mu} + \bar{\mathbf{m}}\right)||_2, \tag{16}$$

where $||\mathbf{v}||_2 = \sqrt{\mathbf{v}'\mathbf{v}}$. A minimum-norm solution can be found using the Moore-Penrose inverse [25], among other possible numerical algorithms [26]. The Moore-Penrose inverse was used in the context of IO analysis by Pereira *et al.* [27].

The determination of posterior covariances, which is mathematically more complex, is described in Appendix A.4.

2.4.2. The WLS Algorithm

The *weighted least-squares (WLS) algorithm* is valid when aggregate data is maximally uninformative, in which case all correlations between disaggregate data are approximately unitary, $r_{jk} = 1$, as shown in Appendix A.5.

Two additional assumptions are now considered: first, that each disaggregate numerical datum is affected by few accounting identities; second, that each accounting identity affects many disaggregate numerical data. That is, the row sums of matrix $\mathbf{G}$ are large integers, while column sums are small (but positive) integers. These auxiliary assumptions are likely to be met in practice.

In Appendix A.7 it is shown that under these conditions the data balancing algorithm is given by:

$$\mathbf{m} = \boldsymbol{\mu} + \hat{\boldsymbol{\sigma}}\mathbf{G}'\boldsymbol{\alpha}, \tag{17}$$

and the Lagrange multipliers are determined by:

$$\left(\mathbf{G}\hat{\boldsymbol{\sigma}}\mathbf{G}'\right)\boldsymbol{\alpha} = -\left(\mathbf{G}\boldsymbol{\mu} + \bar{\mathbf{m}}\right). \tag{18}$$

This is a weighted least-squares (WLS) method in which the weights are prior *uncertainties*.

2.4.3. The Proportional Algorithm

A further simplification is the situation when all (aggregate and disaggregate) prior relative uncertainties are identical, $\sigma_j/\mu_j = \text{const}$. As shown in Appendix A.5, in this case all correlations are unitary, $\rho_{jk} = 1$. This simplification leads to the *proportional algorithm*.

If the same considerations about data structure of the WLS case still apply (many numerical data per accounting identity, few accounting identities per numerical datum), Equation (17) becomes:

$$\mathbf{m} = \boldsymbol{\mu} + \hat{\boldsymbol{\mu}}\mathbf{G}'\boldsymbol{\alpha}. \tag{19}$$

Each row of the previous expression is:

$$m_j = \mu_j \left(1 + \sum_{i=1}^{n_K} G_{ij}\alpha_i\right). \tag{20}$$

Recall that the Taylor first-order approximation of e^x where $x \simeq 0$ is $e^x \simeq 1 + x$. If the update rule is applied recursively in small steps, the previous expression can be rewritten as:

$$m_j = \mu_j \prod_{i=1}^{n_K} (\gamma_i)^{G_{ij}}, \tag{21}$$

where $\gamma_i = e^{\alpha_i}$. The first-order constraint, Equation (8), can be expressed in scalar form as:

$$0 = \sum_{j=1}^{n_T} G_{ij} m_j + \bar{m}_i. \tag{22}$$

Combining the two previous expressions and imposing that the multipliers are adjusted one at a time, by balancing the respective first-order constraint, leads to:

$$0 = \gamma_i \sum_{j=1}^{n_T} G_{ij}^P \mu_j - \frac{1}{\gamma_i} \sum_{j=1}^{n_T} G_{ij}^N \mu_j + \bar{m}_i, \tag{23}$$

where $G_{ij}^P = 1$ if $G_{ij} = 1$ and zero otherwise, and where $G_{ij}^N = 1$ if $G_{ij} = -1$ and zero otherwise. The previous expression yields the solution:

$$\gamma_i = \frac{-\bar{m}_i \pm \sqrt{\bar{m}_i^2 + 4 \left(\sum_{j=1}^{n_T} G_{ij}^P \mu_j \right) \left(\sum_{j=1}^{n_T} G_{ij}^N \mu_j \right)}}{2 \sum_{j=1}^{n_T} G_{ij}^P \mu_j}, \tag{24}$$

if $\sum_{j=1}^{n_T} G_{ij}^P \mu_j > 0$ and otherwise:

$$\gamma_i = \frac{\sum_{j=1}^{n_T} G_{ij}^N \mu_j}{\bar{m}_i}. \tag{25}$$

The algorithm consists in the application of Equation (23) to each accounting identity separately to determine the Lagrange parameters and the update of the best guess estimates by the application of Equation (21). This is a generalization of the popular RAS method for arbitrary structure.

The derivation of the proportional algorithm started by considering that all priors are maximally uninformative. However, the critical assumption is that all prior relative uncertainties are identical, $\sigma_j/\mu_j =$ constant, which implies that all correlations are unitary.

2.4.4. The Linear Algorithm

In the two least-squares methods derived above it is necessary to solve linear systems, by calculating a matrix inverse, a pseudo-inverse or using some implicit method [26]. These are operations of greater complexity than the simple iterative rule of the proportional method. The *linear algorithm* derived now is a variation of the WLS algorithm which does not require solving a linear system but that can take into account uncertainty information, which the proportional algorithm does not.

Consider that each accounting identity, $\mathbf{g}(i)$, is used to determine the corresponding Lagrange multiplier, α_i, in isolation. Thus, Equation (17) becomes: vspce-12pt

$$\mathbf{m}(i) = \boldsymbol{\mu} + \alpha_i \hat{\boldsymbol{\sigma}} \mathbf{g}(i)', \tag{26}$$

and direct substitution in Equation (8) leads to the solution:

$$\alpha_i = -\frac{\mathbf{g}(i)\boldsymbol{\mu} + \bar{m}_i}{\mathbf{g}(i)\hat{\boldsymbol{\sigma}}\mathbf{g}(i)'}. \tag{27}$$

The adjustment induced by the previous expression is linear (as opposed to the multiplicative adjustment of the proportional algorithm) and can be applied simultaneously to all pairs of accounting identities and Lagrange multipliers. The linear algorithm consists in the application of Equation (17) and:

$$\boldsymbol{\alpha} = -\left(\mathbf{G}\boldsymbol{\mu} + \bar{\mathbf{m}}\right) \div \operatorname{diag}\left(\mathbf{G}\hat{\boldsymbol{\sigma}}\mathbf{G}'\right), \tag{28}$$

where $\div$ is Hadamard (or entry-wise) division.

3. Comparison with Conventional Methods

3.1. Proportional and Cross-Entropy Methods

Data balancing occurs in IO analysis under different circumstances, of which the most thoroughly explored is the problem of table update when row and column sums, $\bar{m}_i^R$ and $\bar{m}_j^C$, are known for the current year, and interior points, μ_{ij}^*, are known from a previous year [13]. In this problem, the goal is to update each interior point to m_{ij}^*, such that $\mathbf{M}^*\mathbf{e} = \bar{\mathbf{m}}^{\mathbf{R}}$ and $\mathbf{M}^{*\prime}\mathbf{e} = \bar{\mathbf{m}}^{\mathbf{C}}$. In the previous expressions $\mathbf{e}$ is a vector of ones, $'$ is transpose and superscript * was added to distinguish the original data in dense format from the data in sparse format introduced in the following paragraphs.

The most popular strategy to address this problem is a biproportional method in which the original matrix is iteratively multiplied by a left and a right perturbation diagonal matrices, $m_{ij}^* = \mu_{ij}^*\gamma_i^R\gamma_j^C$ (where the γ's are multiplicative adjustment factors), until the row and column sums are satisfied. The first such technique to be used in IO analysis was the RAS method [28,29], which spawned a vast offspring, whose genealogy is reviewed in Lahr and de Mesnard [5] and whose mathematical properties are characterized in de Mesnard [30].

Although this method is referred to as being specifically biproportional, it can be recast in a more general framework in which the numerical data is affected by any number of constraints, and not necessarily aligned in rows and columns. That is, $m_j = \mu_j \prod_{i=1}^{n_K} \gamma_i^{G_{ij}}$, under the constraint $\mathbf{G}\mathbf{m} = \bar{\mathbf{m}}$, where $\mathbf{m}$ and $\boldsymbol{\mu}$ are now the *prior* and *posterior* vectors, $\bar{\mathbf{m}}$ is a vector of *numerical* constraints and, in the case of a 2×2 matrix:

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}; \quad \bar{\mathbf{m}} = \begin{bmatrix} \bar{m}_1^R \\ \bar{m}_2^R \\ \bar{m}_1^C \\ \bar{m}_2^C \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_{11}^* \\ \mu_{12}^* \\ \mu_{11}^* \\ \mu_{22}^* \end{bmatrix}; \quad \mathbf{m} = \begin{bmatrix} m_{11}^* \\ m_{12}^* \\ m_{11}^* \\ m_{22}^* \end{bmatrix}. \tag{29}$$

(Notice that this is the same problem defined by Equation (4), but now expressed in terms of observable best guesses.)

Cross-entropy methods, such as Snickars and Weibull [31] (see also Golan and Vogel [32], Robinson *et al.* [20] or Fernandez-Vasquez [33]), address the table update problem using a constrained optimization framework, in which the objective function is cross entropy [18]. That is, a Lagrangean is defined as some variation of:

$$L = \sum_{j=1}^{n_T} m_j \log \frac{m_j}{\mu_j} + \sum_{i=1}^{n_K} \lambda_i \left(\sum_{j=1}^{n_K} G_{ij} m_j - \bar{m}_i \right), \tag{30}$$

where the first term in the right hand side is the relative (or cross) entropy to be minimized, followed by the set of constraints.

Cross-entropy minimization provides a posterior *probability distribution* that is closest to the prior in an information sense and is also consistent with the constraints, as discussed in Section 2.2. Therefore, the application of this technique in this context implies that IO quantities, either economic transactions or technical coefficients, are being treated as probabilities, and that the IO table as a whole (either transaction or technical) is viewed as a probability distribution. This interpretation should be contrasted with the Bayesian approach of Section 2.2 in which a numerical datum is represented by a random variable instead of a real number.

Minimization of the Lagrangean with respect to the posteriors yields a solution of the form:

$$t_j = \frac{\theta_j}{Z} \exp \left(\sum_{i=1}^{n_K} \lambda_i G_{ij} \right), \tag{31}$$

where Z is a normalization constant. Substitution of $\gamma_i = \exp(\lambda_i)/Z^{1/n_K}$ leads to the finding, in agreement with Bacharach [34], that the solution of such a problem is none other than the simple RAS method described above. Thus, cross-entropy methods provide a theoretical interpretation of proportional methods in a transaction-as-probability sense. In fact, to describe a method as being proportional or cross-entropy is to view the same object under two different angles: "proportionality" describes the implementation algorithm while "cross entropy" describes the objective function.

An interesting variation of cross entropy is the concept of *generalized* cross entropy of Golan *et al.* [21]. They address the classical problem formulated in the first paragraph of this subsection, but expressed in terms of technical coefficients instead of transaction values. For consistency with the remainder of the present exposition, their problem is now reformulated as determining the interior points of a matrix, m^*_{ij}, subject to fixed row and column constraints, $\bar{m}^R_i$ and $\bar{m}^C_j$, given priors μ^*_{ij}. They introduce a support of M discrete points, $0 \leq q_{ij1} < \ldots < q_{ijM} \leq \bar{m}^C_i$, and a probability associated with each point, $p(q_{ijk})$, so that (according to the Equation (24) in [21]), the datum is actually an expectation:

$$m^*_{ij} = \sum_{k=1}^{M} q_{ijk} p(q_{ijk}). \tag{32}$$

The optimization problem they consider (Equations (20)–(23) in [21]) is:

$$\begin{aligned} L = &\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{M} p(q_{ijk}) \log \left(\frac{p(q_{ijk})}{\pi(q_{ijk})} \right) + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij} \left(1 - \sum_{k=1}^{M} p(q_{ijk}) \right) \\ &+ \sum_{i=1}^{n} \alpha^R_i \left(\bar{m}^R_i - \sum_{j=1}^{n} m^*_{ij} \right) + \sum_{j=1}^{n} \alpha^C_j \left(\bar{m}^C_j - \sum_{j=1}^{n} m^*_{ij} \right). \end{aligned} \tag{33}$$

But this is is none other than a multivariate version of Equation (6) where the zero and first order constraints are known. Thus, in spite of some technical inaccuracies (e.g., Equation (25) in [21] defines

an object which is the variance of an expectation), the generalized cross entropy of Golan *et al.* [21] is a forerunner of the Bayesian theory of IO uncertainty developed here.

3.2. Least-Squares Methods

There are balancing methods that use constrained optimization with other objective functions besides cross entropy [9], of which the most popular are least-squares methods [35–39]. In these studies the Lagrangean is some variation of:

$$L = \frac{1}{2}\left(\mathbf{m}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{m}-\boldsymbol{\mu}\right) + \sum_{i=1}^{n_K}\lambda_i\left(\sum_{j=1}^{n_T}G_{ij}m_j - \bar{m}_i\right). \tag{34}$$

Now the numerical datum is no longer characterized only by a best guess (or expected value) μ_j or m_j, but also by an uncertainty (or standard-deviation), σ_j or s_j. The corresponding covariance matrices are defined as $\boldsymbol{\Sigma} = \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}$ and $\mathbf{S} = \hat{\mathbf{s}}\mathbf{R}\hat{\mathbf{s}}$, where $\mathbf{P}$ and $\mathbf{R}$ are the prior and posterior correlation matrices, and $\hat{}$ denotes diagonal matrix.

In these studies, prior uncertainty is defined as $\sigma_j = a_j\mu_j$, the product of the best guess prior and a *reliability index* a_j, which expresses the subjective degree of belief that the expert has about the accuracy of the data. Some of these studies [38] consider zero prior correlations, leading to a solution of the form:

$$\mathbf{m} = \boldsymbol{\mu} + \hat{\boldsymbol{\sigma}}^2\mathbf{G}'\boldsymbol{\lambda}. \tag{35}$$

Other studies obtain covariances from considerations of time autocorrelation [36,39], leading to a solution:

$$\mathbf{m} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{G}'\boldsymbol{\lambda}. \tag{36}$$

In a variation to this theme, Rampa [10] notes that a second-order Taylor expansion to the cross-entropy objective function is a weighted least square objective function with the weight being the prior best guess:

$$m\log\frac{m}{\mu} \simeq (m-\mu) + \frac{1}{2}\frac{(m-\mu)^2}{\mu} + \ldots \tag{37}$$

With this insight, he proposes a subjective weighted least-squares (SWLS) method where, as before, $\sigma_j = a_j\mu_j$ and a_j is a reliability index, but now the weights in the objective function are not covariances but standard-deviations, leading to a solution of the form:

$$\mathbf{m} = \boldsymbol{\mu} + \hat{\boldsymbol{\sigma}}\mathbf{G}'\boldsymbol{\lambda}. \tag{38}$$

This proposal should be contrasted with the KRAS method [8], which addresses the issue of conflicting constraints, i.e., the table update problem in which the goal is to find $\mathbf{M}^*\mathbf{1} = \bar{\boldsymbol{\mu}}^{\mathbf{R}}$ and $\mathbf{M}^{*\prime}\mathbf{1} = \bar{\boldsymbol{\mu}}^{\mathbf{C}}$ but now the constraints are themselves inconsistent, $\mathbf{1}'\bar{\boldsymbol{\mu}}^{\mathbf{R}} \neq \mathbf{1}'\bar{\boldsymbol{\mu}}^{\mathbf{C}}$. The method assumes that for each *constraint* prior both a best guess, $\bar{\mu}_i$, and an uncertainty, $\bar{\sigma}_j$, are available. The method consists in alternating a proportional adjustment of interior points (conventional RAS) and an adjustment of constraints of the form:

$$\bar{m}_i^R = \bar{\mu}_i^R + \bar{\sigma}_i^R\lambda; \tag{39}$$

$$\bar{m}_i^C = \bar{\mu}_i^C - \bar{\sigma}_i^C\lambda. \tag{40}$$

The previous expressions are a particular case of the SWLS method, with a single accounting identity, $\mathbf{gm} = 0$, which, in the case of a 2×2 matrix, becomes:

$$\mathbf{g} = \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}; \quad \mathbf{m} = \begin{bmatrix} \bar{m}_1^R \\ \bar{m}_2^R \\ \bar{m}_1^C \\ \bar{m}_2^C \end{bmatrix}. \tag{41}$$

So the KRAS method, implementation details aside, is actually a hybrid between a cross-entropy optimization problem (for interior points) and a weighted least-squares optimization problem (for constraints).

For clarity, this example considered row and column sums as constraints, but the KRAS method allows for arbitrary structure (i.e., a numerical constraint can be linked to any subset of disaggregate data). However, the KRAS method always requires the classification of data quality into two quality levels, where data in the first quality level is adjusted using the proportional algorithm and data in the second data level is adjusted using the linear algorithm (which is identical to the SWLS method in the case of a single accounting identity).

The hybrid character shares some affinities with the work of Lieu *et al.* [40] and Lieu and Hicks [41], who combine entropy maximization and least-squares minimization to address conflicting constraints.

3.3. Discussion

The conventional methods, reviewed in Sections 3.1 and 3.2 are not very useful for the general data balancing problem outlined in Section 2.1 due to several problems, all of which are solved under the Bayesian approach.

Objective function: Conventional data balancing methods can be formulated as constrained optimization problems with different objective functions: cross entropy, generalized least squares, least squares weighted with variances or least squares weighted with standard-deviations. However, they offer no obvious rule to determine when each method should be applied under a particular circumstance.

Fortunately, the conventional methods reviewed here are very similar to the algorithms derived in Section 2.4, which means that it is possible to identify when each conventional method is valid. The GLS method should be used when uncertainty estimates for some disaggregate data and all aggregate data are available. The WLS method should be used when uncertainty estimates for some disaggregate data only are available. The proportional method should be used when no uncertainty data is available. The Bayesian algorithms are all iterative while conventional least-squares methods take place in a single step, which means that the latter are only valid if the initial inconsistency is small.

Uncertainty estimates: All data balancing methods can use information on best guesses, but they differ in the ability to incorporate information on uncertainty. In cross-entropy/biproportional methods there is no obvious way to introduce such information (but it can be done, e.g., [42]), while in least-square methods it is mandatory to specify both the standard-deviation and the correlations of the prior using subjective reliability indices which require expert knowledge of the data.

These problems do not occur in the methods proposed here, which adhere strictly to the rule that only available information should be used, meaning that if some uncertainty or correlation is missing

the worst-case scenario must be assumed: the uncertainty equals the best guess and the correlation is unitary. This strategy provides objective rules that can be used even in the absence of expert knowledge.

Data quality: Another important characteristic of conventional methods is that either some part of the numerical data is adjusted while the other is held fixed (cross-entropy and proportional methods), or all the data is adjusted at the same time (least-squares methods).

In a Bayesian context, it is possible to introduce a ranking of data quality in the sequence in which the balancing procedure is implemented, as described in Section 2.3. In practice, a qualitative ordering of numerical data by level of trustworthiness is often more accessible than quantitative uncertainty estimates. Unlike conventional methods, the Bayesian approach to data balancing allows the user to make direct use of this knowledge.

There are two additional problems that some, but not all conventional methods exhibit, and which are absent from the algorithms derived here.

Arbitrary structure: Proportional methods assume that the data is organized in a matrix format [43] while cross-entropy and least-squares methods allow for an arbitrary structure.

Sign preservation: Cross-entropy and proportional methods always ensure sign preservation, while least-squares methods do not (i.e., an initially positive datum may become negative). In practice, all transactions in a table should be positive, and balancing items such as fixed capital formation, variations in stocks or net taxes can take both signs. Appendix A.1 shows how to allow balancing items to shift sign while ensuring the sign preservation of transactions.

In summary, this work has achieved a major theoretical unification in the problem of data balancing by being able to state the conditions in which conventional methods are valid and by providing simple rules to determine missing second-order data. The range of source information that can be used in the data balancing problem was expanded (data quality) while interesting features that some conventional methods possess have been kept (arbitrary structure and sign preservation).

3.4. Empirical Considerations

Section 2.4 proposed a series of data balancing algorithms, whose choice depends on the available information on source uncertainties and correlations. This subsection presents a brief survey of the empirical literature and discuss its implication for the choice of algorithm.

The Bayesian algorithm that requires more detailed source data is the GLS method, so the review starts by the least-squares literature, in which relative uncertainties are referred to as reliability indices. Weale [38] considers three reliability indices: 1.5%, 6.5% and 15%; Byron *et al.* [44] consider four: 3%, 13%, 30% and 50%; Chen [45] considers three: 10%, 20% and 30%. Using a two-step procedure that involves determining first a qualitative reliability indicator and later a coefficient of variation, Rassier *et al.* [46] consider relative uncertainties that range from 0% to 100%. Rampa [10] considers five: 1, 1.5, 2, 3 and 4 times the smallest value (in least-squares methods the absolute value of the reliability index is not important, only the relative value). Some studies [38,39] go as far as estimating prior covariances from time auto-correlations, but they are routinely assumed to be zero.

From this very brief survey it is apparent that in least-squares methods the reliability indices do not express factual quantitative knowledge but only a broad sense of qualitative ordering of different types of

data (row/column sums, domestic transactions, added value, imports, *etc.*). In the Bayesian framework this qualitative ordering should be used directly in the form of data quality. The assignment of subjective reliability indices in the absence of numerical empirical support is not only unnecessary but also contrary to the Bayesian philosophy, according to which only available information should be used in the data balancing problem.

Uncertainty estimates are not routinely provided by statistical offices, but such studies are occasionally produced, and their results are broadly consistent, indicating that relative uncertainty, σ_j/μ_j, of IO transactions decreases monotonically with the best guess, μ_j, in the broad range of 40% to 10% and row/column sums are known with proportionately better accuracy than interior points [47], decreasing down to 3%. These broad trends have been confirmed by studies in different countries, such as Bullard and Sebald [48] for the USA, Lenzen [49] for Australia, Nhambiú [50] for Portugal, and Lenzen *et al.* [51] for the UK. Yamakawa and Peters [52] use time-series inconsistencies to calculate source data uncertainty and Díaz and Morillas [53] use fuzzy logic [54,55] and firm-level data to estimate the uncertainty of technical coefficients.

The detailed studies of source uncertainty mentioned in the preceding paragraph are very labour intensive, so in a study that involves gathering empirical data of this type it probably makes sense to use the GLS data balancing method of Section 2.4. However, if uncertainty estimates are obtained from a literature survey, it may be better to use the WLS algorithm (or its linear variant) and to assign a higher quality level to aggregate data. The computational effort of using the full GLS method is several orders of magnitude higher and, in the absence of a high degree of confidence in the quality of source data, that additional effort is unjustified.

In conclusion, this paper suggests the following: if no quantitative uncertainty estimates are available, use only knowledge of the ranking of data quality. Unless there is substantial confidence in the uncertainty estimates of aggregate data, use a simplification instead of the full GLS method. The linear algorithm is the most flexible and easiest to implement and its theoretical shortcomings are probably of no consequence for most empirical applications.

4. Conclusions

This paper studies the problem of IO data balancing from the standpoint of Bayesian inference.

The basic idea that motivates the present work is very simple, although its implications are far from trivial. A numerical datum known with some degree of uncertainty is treated as a random variable, t_j, whose probability density function, $p_j(q)$, quantifies the degree of belief that the datum takes realization q_j. The numerical datum is characterized empirically by a best guess, m_j, and an uncertainty, s_j, which are interpreted as the expectation and standard deviation of random variable t_j.

The set of posteriors, $\mathbf{t}$, must satisfy a set of accounting identities, summarized as $\mathbf{G}\mathbf{t} = \mathbf{0}$, and a set of priors, $\boldsymbol{\theta}$ is initially available, such that $\mathbf{G}\boldsymbol{\theta} \neq \mathbf{0}$. Application of the cross-entropy minimization, subject to first and second moment constraints leads to the analytical solution of the data balancing problem.

Several numerical algorithms are derived, whose scope of application depends on the availability of uncertainty estimates. If no uncertainty information is available, the algorithm is a natural generalization of the familiar RAS method. If some uncertainty estimates are available but there is no guarantee

that the uncertainty of aggregate data was obtained independently from that of disaggregate data, then the algorithm is an uncertainty-weighted least-squares method. If the uncertainty of both aggregate and disaggregate data was obtained independently, then the algorithm is an alternate generalized least-squares method for first and second moment parameters. All algorithms are iterative and valid for arbitrary structure.

This paper presents a review of conventional data balancing algorithms and establishes a one-to-one correspondence with the Bayesian algorithms derived earlier, thus underpinning the assumptions of each conventional method. In particular, this paper finds that the conventional RAS method is a particular case of the proportional algorithm and thus implicitly assumes that the relative uncertainty of all data points is identical.

This paper's suggestion for practical implementation (in the absence of high-quality uncertainty data) is the use of the linear algorithm described in Section 2.4, combined with the assignment of both uncertainty estimates (when available) and of data quality to the numerical priors.

Acknowledgments

This paper benefitted from the comments of many colleagues and friends. In particular, I want to thank the suggestions and references provided by Michael Lahr, Esteban Fernandez-Vasquez and Umed Temurshoev. I also want to thank my wife Kamila and my friend Natalia for correcting the English language in the manuscript. Any errors it may contain are the author's sole responsibility. Finally, I want to acknowledge the financial support of the Portuguese Science and Technology Foundation FCT), in the aim of project MeSur (reference: PTDC/SEN-ENR/111710/2009).

Conflicts of Interest

The author declare no conflicts of interest.

A. Appendix

A.1. Treatment of Zero and Negative Entries

Sometimes IO tables report negative entries and they usually report many zeros, while in the Bayesian data balancing method it is assumed that all IO data is positive. These situations are handled as follows.

Negative entries can appear in an IO table by convention: subsidies are a negative primary factor; net exports can be negative (if imports exceed exports); *etc.* In symmetric product-by-product IO tables derived using the product technology negative entries can appear too [56].

The present paper is not concerned with the question of whether negative entries are meaningful or not, but with the question of how to handle them, if the practitioner believes that they are.

Thus, for the purpose at hand, the important question is whether the entry should be allowed to shift sign during the balancing procedure (e.g., if it is a change in stocks) or not (e.g., a trade margin in a SAM).

A negative entry that is not allowed to shift sign can be handled by altering the structure of the problem. Consider the problem defined by Equation (5) where entry Z_{12} is negative. The acccounting identities of the original system are:

$$Z_{11} + Z_{12} = x_1^R; \tag{42}$$
$$Z_{21} + Z_{22} = x_2^R; \tag{43}$$
$$Z_{11} + Z_{21} = x_1^C; \tag{44}$$
$$Z_{21} + Z_{22} = x_2^C. \tag{45}$$

This problem can be recast as:

$$Z_{11} = x_1^{R*}; \tag{46}$$
$$Z_{21} + Z_{22} = x_2^R; \tag{47}$$
$$Z_{11} + Z_{21} = x_1^C; \tag{48}$$
$$Z_{22} = x_2^{C*}; \tag{49}$$
$$x_1^{R*} = (-Z_{12}) + x_1^R; \tag{50}$$
$$x_2^{C*} = (-Z_{12}) + x_2^C, \tag{51}$$

where all quantities (including $-Z_{12}$) are now positive numbers.

To allow an IO datum to shift sign, is is necessary to consider two different positive-valued entries in an IO table: the superavit component, an entry in the corresponding row, and the deficit component, in the corresponding column. If, for example, the datum is positive, then it is assigned to the superavit component, and an infinitesimal is assigned to the deficit component.

To allow a balancing item to shift sign using the linear algorithm (which in this paper is recommended for practical uses), it is sufficient not to enforce the displacement bound. That is, if t_j is a transaction, it is necessary to ensure that at every step $|1 - m_j/\mu_j| < \epsilon$. If t_j is a balancing item that check should not be performed.

A zero value in an IO table can mean two different things: either the transaction is logically impossible or it was simply too small to have been recorded. In the first case, it should be excluded from the set of numerical data.

However, if a transaction t_j is below the resolution of the IO table, ϵ, but it is nonetheless logically possible, it should be assigned a maximally uninformative prior, $\sigma_j = \mu_j$, with an infinitesimal best guess, $\mu_j \ll \epsilon$.

Because the initial inconsistency of IO tables is usually small, this step is unnecessary, i.e., $m_j \ll \epsilon$ so there is no problem in removing transaction t_j from the set of numerical data altogether.

This paper suggests the explicit consideration of infinitesimals only when there is the suspicion that some non-infinitesimal best guess is misreported, a situation studied by Keogh and Quill [4].

A.2. Stochastic First and Second Moment Constraints

The first moment constraint of Equation (1) reads:

$$0 = \mathrm{E}\left[\sum_{j=1}^{n_T} G_{ij} t_j + k_i\right] = \sum_{j=1}^{n_T} G_{ij}\,\mathrm{E}[t_j] + \mathrm{E}[k_i], \tag{52}$$

and if $\bar{m}_i = \mathrm{E}[k_i]$, the previous expression becomes:

$$0 = \sum_{j=1}^{n_T} G_{ij} m_j + \bar{m}_i. \tag{53}$$

In matrix form, this equation leads to Equation (8).

The second moment constraint of Equation (1) is a bit more problematic, because of correlations. Consider that $k_i < 0$ and that $G_{ij} = 1$ if $1 \leq j \leq n_P$ while $G_{ij} = -1$ if $n_P + 1 \leq j \leq n_T$. Consider that Equation (1) is rearranged as:

$$-k_i - \sum_{j=n_P+1}^{n_T} G_{ij} t_j = \sum_{j=1}^{n_P} G_{ij} t_j. \tag{54}$$

The previous expression contains a sum of positively valued random variables on the left hand side and another such sum on the right hand side. Application of second moments leads to:

$$\mathrm{Var}\left[-k_i - \sum_{j=n_P+1}^{n_T} G_{ij} t_j\right] = \mathrm{Var}\left[\sum_{j=1}^{n_P} G_{ij} t_j\right]. \tag{55}$$

From basic probability theory:

$$\mathrm{Var}\left[\sum_{i=1}^{n} t_i\right] = \sum_{i=1}^{n} \mathrm{Var}\left[t_i\right] + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathrm{Cov}[t_i, t_j], \tag{56}$$

where $\mathrm{Cov}[t_i, t_j] = r_{ij} s_i s_j$. If the numerical constraint is uncorrelated with the numerical data, the application of this rule to Equation (55) leads to:

$$\begin{aligned}\mathrm{Var}[k_i] + \sum_{j=n_P+1}^{n_T} G_{ij}^2\,\mathrm{Var}[t_j] + 2\sum_{j=n_P+1}^{n_T}\sum_{k=n_P+1}^{j-1} G_{ij} G_{ik}\,\mathrm{Cov}[t_j, t_k] = \\ \sum_{j=1}^{n_P} G_{ij}^2\,\mathrm{Var}[t_j] + 2\sum_{j=1}^{n_P}\sum_{k=1}^{j-1} G_{ij} G_{ik}\,\mathrm{Cov}[t_j, t_k].\end{aligned} \tag{57}$$

With the substitution $\bar{s}_i^2 = \mathrm{Var}[k_i]$, the previous expression becomes:

$$\begin{aligned}\bar{s}_i^2 + \sum_{j=n_P+1}^{n_T} G_{ij}^2 s_j^2 + 2\sum_{j=n_P+1}^{n_T}\sum_{k=n_P+1}^{j-1} G_{ij} G_{ik} r_{jk} s_j s_k = \\ \sum_{j=1}^{n_P} G_{ij}^2 s_j^2 + 2\sum_{j=1}^{n_P}\sum_{k=1}^{j-1} G_{ij} G_{ik} r_{jk} s_j s_k.\end{aligned} \tag{58}$$

Notice that r_{jk} when $G_{ij} = 1$ and $G_{ij} = -1$ are absent from Equation (58).

The matrix form of Equation (58) is given by Equation (9). Notice that in Equation (9) the concordance matrix appears two times, and one of those times it is in absolute terms, $|\cdot|$ (although it does not matter which one). This is necessary so that correlations between r_{jk} when $G_{ij} = 1$ and $G_{ij} = -1$ cancel out.

A.3. Analytical Solution

The information about the first two moments is introduced in the Lagrangean of the system in scalar form as:

$$\begin{aligned} L = & \int_{\Omega} dq\, p(\mathbf{q}) \ln\left(\frac{p(\mathbf{q})}{\pi(\mathbf{q})}\right) + \lambda\left(\int_{\Omega} dq\, p(\mathbf{q}) - 1\right) \\ & + \sum_{i=1}^{n_K} \alpha_i^* \left(\sum_{j=1}^{n_T} G_{ij} \int_{\Omega} dq\, p(\mathbf{q}) q_j + \bar{m}_i\right) \\ & + \sum_{i=1}^{n_K} \beta_i^* \left(\sum_{j=1}^{n_T} G_{ij} \left(\left(\int_{\Omega} dq\, p(\mathbf{q}) q_j^2\right) - m_j^2\right)\right. \\ & \left. + 2 \sum_{j=1}^{n_T} \sum_{k=1}^{j-1} G_{ijk}^* \left(\left(\int_{\Omega} dq\, p(\mathbf{q}) q_j q_k\right) - m_j m_k\right) + \bar{s}_i^2\right). \end{aligned} \tag{59}$$

In Equation (59) the expression $\int_{\Omega} dq$ is a shorthand for the product $\prod_{j=1}^{n_T} \int_0^{\infty} dq_j$. Each q_j is the realization of the random variables t_j and θ_j. The first term on the right hand side of Equation (59) contains the entropy of the posterior, relative to the prior. The second term is the normalization constraint. The third term is the set of best guess constraints. The fourth term is the set of uncertainty constraints. The term m_j is the marginal expectation of t_j, defined as:

$$m_j = \int_{\Omega} dq\, p(\mathbf{q}) q_j. \tag{60}$$

The term G_{ijk}^* is $G_{ijk}^* = 1$ if $G_{ij} = G_{ik}$ for some i, j and $k \neq j$, or $G_{ijk}^* = 0$ otherwise. The λ, α^*'s and β^*'s are, respectively, the Lagrange multipliers of the normalization, best guess and uncertainty constraints. Minimization of Equation (59) with respect to $p(\mathbf{q})$ yields:

$$\begin{aligned} 0 = & -(\ln p(\mathbf{q}) + 1)\frac{1}{\ln \pi(\mathbf{q})} + \lambda + \sum_{j=1}^{n_T} \left(\sum_{i=1}^{n_K} G_{ij} \alpha_i^*\right) q_j \\ & + \sum_{j=1}^{n_T} \left(\left(\sum_{i=1}^{n_K} G_{ij} \beta_i^*\right)(q_j^2 - 2 q_j m_j)\right) \\ & + \sum_{j=1}^{n_T} \sum_{k=1}^{j-1} \left(2\left(\sum_{i=1}^{n_K} G_{ijk}^* \beta_i^*\right)(q_j q_k - q_j m_k)\right) + C. \end{aligned} \tag{61}$$

The C's in the previous and subsequent expressions denote appropriately chosen constants. The previous expression can be rewritten in the form:

$$p(\mathbf{q}) = \pi(\mathbf{q})C \exp\left(\sum_{j=1}^{n_T}\left(\sum_{i=1}^{n_K} G_{ij}\beta_i^*\right)q_j^2 + \sum_{j=1}^{n_T}\sum_{k=1}^{j-1} 2\left(\sum_{i=1}^{n_K} G_{ijk}^*\beta_i^*\right)q_j q_k + \right. \tag{62}$$

$$\left. + \sum_{j=1}^{n_T}\left(\sum_{i=1}^{n_K} G_{ij}\alpha_i^* - 2\left(m_j\sum_{i=1}^{n_K} G_{ij}\beta_i^* + \sum_{k=1}^{n_T} m_k\left(\sum_{i=1}^{n_K} G_{ijk}^*\beta_i^*\right)\right)\right)q_j\right). \tag{63}$$

This expression can be simplified to:

$$p(\mathbf{q}) = \pi(\mathbf{q})C \exp\left(-\sum_{j=1}^{n_T}\left(\sum_{i=1}^{n_K} G_{ij}\frac{\beta_i}{2}\right)q_j^2 - \sum_{j=1}^{n_T}\sum_{k=1}^{j-1} 2\left(\sum_{i=1}^{n_K} G_{ijk}^*\frac{\beta_i}{2}\right)q_j q_k + \right.$$

$$\left. + \sum_{j=1}^{n_T}\left(\sum_{i=1}^{n_K} G_{ij}\alpha_i\right)q_j\right), \tag{64}$$

where $\beta_i = -\beta_i^*/2$ and $\alpha_i = \alpha_i^* - 2\beta_i^* m_i^*$, and the latter term is obtained as:

$$m_j\sum_{i=1}^{n_K} G_{ij}\beta_i^* + \sum_{k=1}^{n_T} m_k\left(\sum_{i=1}^{n_K} G_{ijk}^*\beta_i^*\right) =$$

$$\sum_{i=1}^{n_K}\beta_i^*\left(G_{ij}m_j + \sum_{k=1}^{n_T} G_{ijk}^* m_k\right) = \sum_{i=1}^{n_K}\beta_i^* m_i^*. \tag{65}$$

Thus, although m_i^* is still unknown, it is only a function of the accounting identity iterator i and independent of the iterators of numerical data j or k. Since the Lagrange multipliers are still free, the substitution above is valid.

Notice that the exponent in Equation (64) is a polynomial whose coefficients are linear combinations of Lagrange multipliers. If the prior is a multivariate truncated Gaussian and the constraints are of second order, the posterior is also a truncated multivariate Gaussian whose probability density is:

$$p(\mathbf{q}) = C \exp\left(-\frac{1}{2}(\mathbf{q} - \tilde{\mathbf{m}})'\tilde{\mathbf{S}}^{-1}(\mathbf{q} - \tilde{\mathbf{m}})\right). \tag{66}$$

The exponent of the prior and posterior probability densities can be expanded in a polynomial form. In particular, Equation (66) becomes:

$$p(\mathbf{q}) = C_1 \exp\left(-\sum_{j=1}^{n_T}\frac{\tilde{s}_{jj}^{-1}}{2}q_j^2 - 2\sum_{j=1}^{n_T}\sum_{k=1}^{j-1}\frac{\tilde{s}_{jk}^{-1}}{2}q_j q_k \right.$$

$$\left. + 2\sum_{j=1}^{n_T}\left(\sum_{k=1}^{n_T}\frac{\tilde{s}_{jk}^{-1}}{2}\tilde{m}_k\right)q_j + C_2\right), \tag{67}$$

and the polynomial expansion of the prior distribution displays a similar pattern. In the previous expression $\tilde{s}_{jk}^{-1}$ is the (j,k) entry of matrix $\tilde{\mathbf{S}}^{-1}$. An explicit expression for the parameters of the posterior

can be obtained by solving expressions of the form $C_{\text{post}} = C_{\text{prior}} + C_{\text{constraint}}$, where each constant is the coefficient of the corresponding polynomial expansion for the posterior and prior distributions and the expressions containing the Lagrange multipliers that result from differentiating the Lagrangean, Equation (59). This leads to Equations (10) and (11).

A.4. GLS Algorithm: Covariances

To determine posterior uncertainties and correlations it is more convenient to express the covariance matrices as $\mathbf{S} = \hat{\mathbf{s}}\mathbf{R}\hat{\mathbf{s}}$ and $\mathbf{\Sigma} = \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}$, so that Equation (10) can be recast as:

$$\hat{\mathbf{s}}^{-1}\mathbf{R}^{-1}\hat{\mathbf{s}}^{-1} = \hat{\boldsymbol{\sigma}}^{-1}\mathbf{P}^{-1}\hat{\boldsymbol{\sigma}}^{-1} + \mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|, \tag{68}$$

where the truncated Gaussian parameters were replaced by observable parameters. Under the substitution $\hat{\boldsymbol{\sigma}}\mathbf{R}^{*}\hat{\boldsymbol{\sigma}} = \hat{\mathbf{s}}\mathbf{R}\hat{\mathbf{s}}$ and $\mathbf{B} = \mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|$, the previous expression simplifies to:

$$\mathbf{R}^{*} = \left(\mathbf{P}^{-1} + \hat{\boldsymbol{\sigma}}\mathbf{B}\hat{\boldsymbol{\sigma}}\right)^{-1}. \tag{69}$$

Using the Woodbury identity [57], the previous expression is equivalent to:

$$\mathbf{R}^{*} = \mathbf{P} - \mathbf{P}\hat{\boldsymbol{\sigma}}\left(\mathbf{B}^{-1} + \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}\right)^{-1}\hat{\boldsymbol{\sigma}}\mathbf{P}. \tag{70}$$

Another application of the Woodbury identity leads to:

$$\mathbf{R}^{*} = \mathbf{P} - \mathbf{P}\hat{\boldsymbol{\sigma}}\left(\mathbf{B} - \mathbf{B}\hat{\boldsymbol{\sigma}}\left(\mathbf{P}^{-1} + \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}\right)^{-1}\hat{\boldsymbol{\sigma}}\mathbf{B}\right)\hat{\boldsymbol{\sigma}}\mathbf{P}. \tag{71}$$

The previous expression can be rewritten as:

$$\mathbf{R}^{*} = \mathbf{P} - \mathbf{P}\hat{\boldsymbol{\sigma}}\mathbf{B}\hat{\boldsymbol{\sigma}}\mathbf{P} + \mathbf{P}\hat{\boldsymbol{\sigma}}\mathbf{B}\hat{\boldsymbol{\sigma}}\left(\mathbf{P}^{-1} + \hat{\boldsymbol{\sigma}}\mathbf{P}\hat{\boldsymbol{\sigma}}\right)^{-1}\hat{\boldsymbol{\sigma}}\mathbf{B}\hat{\boldsymbol{\sigma}}\mathbf{P}. \tag{72}$$

If the displacement from prior to posterior is small, each Lagrange parameter, β_i, is also small. The third term in the right hand side of the previous expression contains products $\beta_i\beta_j \simeq 0$ which can be discarded, and so the first-order approximation is obtained:

$$\mathbf{R}^{*} = \mathbf{P} - \mathbf{F}\#\left(\mathbf{P}\hat{\boldsymbol{\sigma}}\mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|\hat{\boldsymbol{\sigma}}\mathbf{P}\right). \tag{73}$$

The *filter* matrix, $\mathbf{F}$, possesses entry $F_{jk} = 1$ if there is some accounting identity i for which $G_{ij} = G_{ik}$ and $F_{jk} = 0$ otherwise. Matrix $\mathbf{F}$ is introduced to avoid the appearance of mathematical artifacts. After all, Equation (73) is an approximation and, if unchecked, it may lead to the appearance of spurious correlations for an entry (j, k) for which $F_{jk} = 0$. The magnitude of these spurious correlation would be small, but incorrect nonetheless, and with the use of the filter matrix this matter is swiftly addressed.

The Lagrange multipliers are determined by substitution of Equation (73) in Equation (9), leading to:

$$\mathbf{0} = \operatorname{diag}\left(\mathbf{G}\mathbf{\Sigma}|\mathbf{G}|' - \mathbf{G}\mathbf{\Sigma}\mathbf{G}'\hat{\boldsymbol{\beta}}|\mathbf{G}|\mathbf{\Sigma}|\mathbf{G}|'\right) + \bar{\mathbf{s}}^{\mathbf{2}}. \tag{74}$$

(Matrix $\mathbf{F}$ was ignored for computational purposes.) The previous expression can be further simplified by noting that $\operatorname{diag}(\mathbf{A} + \mathbf{B}) = \operatorname{diag}(\mathbf{A}) + \operatorname{diag}(\mathbf{B})$ and that $\mathbf{d} = \operatorname{diag}(\mathbf{A}\hat{\mathbf{b}}\mathbf{C}) = (\mathbf{A}\#\mathbf{C}')\mathbf{b}$, where $\#$ is the Hadamard (or entry-wise) product, since $d_i = \sum_j A_{ij}b_jC_{ji} = \sum_j (A_{ij}C'_{ij})b_j$. This implies that the solution is:

$$\left(\left(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}'\right)\#\left(|\mathbf{G}|\boldsymbol{\Sigma}|\mathbf{G}|'\right)\right)\boldsymbol{\beta}=\operatorname{diag}\left(\mathbf{G}\boldsymbol{\Sigma}|\mathbf{G}|'\right)+\bar{\mathbf{s}}^{\mathbf{2}}. \tag{75}$$

After Equations (73) and (75) have been solved, the posterior uncertainties and correlations are obtained as:

$$s_j=\sigma_j\sqrt{r^*_{jj}}\quad\text{and}\quad r_{jk}=\frac{r^*_{jk}}{\sqrt{r^*_{jj}r^*_{kk}}}. \tag{76}$$

Equations (14), (15), (73) and (75) define a numerical approximation of the analytical solution. The following algorithm is suggested.

First, best guess data are held fixed, while uncertainties and correlations are adjusted, in small steps. That is, Equations (73)–(75) are applied recursively in such a way that the displacement is always small, $|s_j-\sigma_j|/\sigma_j<\epsilon$ and that the solutions are always meaningful, $0<s_j\leq\mu_j$ and $-1\leq r_{jk}\leq 1$. The choice of the convergence parameter ϵ should result from a compromise between computational time and accuracy.

When consistent uncertainties and correlations have been obtained, they are in turn held fixed and best guesses adjusted using Equations (14) and (15). Again, these expressions should be applied in such a way that the displacement is small, $|m_j-\mu_j|/\mu_j<\epsilon$ and the solution is meaningful, $m_j>0$. After every adjustment of the best guess, relative uncertainty is kept fixed, *i.e.*, s_j is replaced by s_jm_j/μ_j and the second-order data are balanced again.

Consider that uncertainty estimates are initially available but correlations are not (and so are assumed to be unitary). In this case, it is natural to consider that uncertainties remain fixed while correlations are adjusted. This idea can be formalized using the filter matrix, $\mathbf{F}$. If the main diagonal is set to zero, $F_{jj}=0$, then $s_j=\sigma_j$ while $r_{jk}\neq\rho_{jk}$. However, it may occur that no solution exists for fixed uncertainties, in which case all second order data must be adjusted, by setting $F_{jj}=1$, thus allowing $s_j\neq\sigma_j$.

A.5. Maximally Uninformative Aggregate Data

For clarity, consider the case of a single accounting identity, indexed with 0, and n disaggregate data, $i=1,\ldots,n$, and consider that both first and second moment constraints are balanced:

$$m_0=\sum_{i=1}^{n}m_i; \tag{77}$$

$$s_0^2=\sum_{i=1}^{n}s_i^2+2\sum_{i=2}^{n}\sum_{j=1}^{i-1}r_{ij}s_is_j, \tag{78}$$

where m_i and s_i are the best guess and absolute uncertainty, and r_{ij} is the correlation. The second expression can be rearranged as:

$$s_0=\sqrt{\sum_{i=1}^{n}s_i^2+2\sum_{i=2}^{n}\sum_{j=1}^{i-1}r_{ij}s_is_j}. \tag{79}$$

Hence, $\partial u_0/\partial r_{ij}>0$, the aggregate uncertainty is a monotonic function of the correlation between disaggregate data. It follows that for aggregate uncertainty to be maximal, then all disaggregate correlations must be maximal too, $r_{ij}=1$.

If all relative uncertainties are identical, $s_0/m_0 = s_j/m_j = \text{const}$, the second-order constraint becomes:

$$m_0^2 = \sum_{i=1}^{n} m_i^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij} m_i m_j, \tag{80}$$

which combined with the first-order constraint leads to:

$$\sum_{i=1}^{n} m_i^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} m_i m_j = \sum_{i=1}^{n} m_i^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij} m_i m_j, \tag{81}$$

which in turn implies:

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} m_i m_j = \sum_{i=2}^{n}\sum_{j=1}^{i-1} r_{ij} m_i m_j, \tag{82}$$

so that $r_{ij} = 1$, all correlations are unitary.

A reviewer of a previous version of this paper stated that unitary correlations are problematic, because the theoretical solution involves matrix inversion and the inverse of a fully correlated covariance matrix is non-invertible. Appendix A.6 shows that the theory has no problem in handling a unitary correlation matrix.

A.6. A Single Accounting Identity

The analytical solution of the covariance update rule, Equation (10), in the case of a single accounting identity with unitary prior correlations, is now studied. It is considered that there are n disaggregate data (labeled from 1 to n) and an aggregate datum (labeled 0).

The first case considered is that in which the aggregate datum and the disaggregate data have the same quality level, so they are adjusted simultaneously. For clarity consider the case $n = 2$. The non-truncated version of Equation (10) is:

$$\hat{\mathbf{s}}^{-1}\begin{bmatrix} \frac{1}{1-r^2} & \frac{-r}{1-r^2} & 0 \\ \frac{-r}{1-r^2} & \frac{1}{1-r^2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\hat{\mathbf{s}}^{-1} = \hat{\boldsymbol{\sigma}}^{-1}\begin{bmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} & 0 \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 \\ 0 & 0 & 1 \end{bmatrix}\hat{\boldsymbol{\sigma}}^{-1} + \begin{bmatrix} \beta & \beta & 0 \\ \beta & \beta & 0 \\ 0 & 0 & -\beta \end{bmatrix}. \tag{83}$$

If the prior correlation is unitary, $\rho = 1$, there seems to be a problem, because the matrix inverse is ill-defined. Does this mean that the Bayesian data balancing method is inconsistent? No. It means that, as Jaynes [6] points out, direct reasoning in terms of infinite quantities (or in this case infinitesimals) should be avoided, since it may lead to paradoxes.

Instead, this paper follows his strict finite-sets policy: "Apply the ordinary processes of arithmetic and analysis only to expressions with a finite number n of terms. Then after the calculation is done, observe how the resulting finite expressions behave as the parameter n increases indefinitely" [6] (p. 452). In the present case, the previous expression can be rewritten as:

$$\hat{\mathbf{s}}^{-1}\begin{bmatrix} 1 & -r & 0 \\ -r & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\hat{\mathbf{s}}^{-1} = \hat{\boldsymbol{\sigma}}^{-1}\begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\hat{\boldsymbol{\sigma}}^{-1} + \begin{bmatrix} \beta^* & \beta^* & 0 \\ \beta^* & \beta^* & 0 \\ 0 & 0 & -\beta \end{bmatrix}. \tag{84}$$

The substitution $\beta^* = (1-\rho^2)\beta$ was performed and it is assumed that $r \simeq \rho$. It is now clear that, as $\rho \to 1$, the Lagrange multiplier affecting disaggregate data vanishes, $\beta^* \ll \beta$. This means that the adjustment effort falls entirely on the aggregate uncertainty, so $s_0 = \sum_{j=1}^{n} \sigma_j$, while $s_j = \sigma_j$ and $r_{jk} = \rho_{jk} = 1$.

The case $n = 2$ was analyzed, but the same result holds in general. Entry (jk) of the inverse of a $n \times n$ matrix $\mathbf{P}$ is:

$$\rho_{jk}^{-1} = \frac{g_{jk}(\mathbf{P})}{\det(\mathbf{P})}, \tag{85}$$

where g_{jk} is potentially a function of every element of $\mathbf{P}$ and $\det$ is the determinant. As in the 2×2 case:

$$\frac{g_{jk}(\mathbf{R})}{s_j s_k} = \frac{g_{jk}(\mathbf{P})}{\sigma_j \sigma_k} + \beta^*; \tag{86}$$

$$\frac{1}{s_0^2} = \frac{1}{\sigma_0^2} - \beta, \tag{87}$$

where $\beta^* = \det(\mathbf{P})\beta$. Since $\beta^* \to 0$ when $\rho \to 1$, all the adjustment effort falls on the aggregate uncertainty, just like in the bivariate case.

The empirically relevant case in which uncertainties estimates are known with better accuracy than correlations is now addressed. In this case $\mathbf{s} = \boldsymbol{\sigma}$ and only correlations are adjusted. If $n = 2$, the non-truncated version of Equation (10) is:

$$\hat{\mathbf{s}}^{-1} \begin{bmatrix} \frac{1}{1-r^2} & \frac{-r}{1-r^2} & 0 \\ \frac{-r}{1-r^2} & \frac{1}{1-r^2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \hat{\mathbf{s}}^{-1} = \hat{\mathbf{s}}^{-1} \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{-\rho}{1-\rho^2} & 0 \\ \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \hat{\mathbf{s}}^{-1} + \begin{bmatrix} 0 & \beta & 0 \\ \beta & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{88}$$

Proceeding as before leads to:

$$\begin{bmatrix} 1 & -r & 0 \\ -r & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\rho & 0 \\ -\rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \hat{\mathbf{s}} \begin{bmatrix} 0 & \beta^* & 0 \\ \beta^* & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \hat{\mathbf{s}}, \tag{89}$$

where $\beta^* = (1-\rho^2)\beta$. There is no indeterminacy in the expression linking prior and posterior correlations:

$$-r = -\rho + s_1 s_2 \beta^*. \tag{90}$$

For arbitrary n the equivalent implicit expression is valid:

$$g_{jk}(\mathbf{R}) = g_{jk}(\mathbf{P}) + s_j s_k \beta^*, \tag{91}$$

where $\beta^* = \det(\mathbf{P})\beta$.

So even in the limit case of a single accounting identity the Bayesian data balancing method generates meaningful results.

A.7. Derivation of the WLS Algorithm

Consider the particular example of a dense IO matrix, where every interior point (ij) is affected by two accounting identities, corresponding to the row and column sums. The expansion of the term $\mathbf{G}'\boldsymbol{\alpha}$ in Equation (14) becomes a vector where each entry is the sum of two Lagrange multipliers, $\alpha_i^R + \alpha_j^C$, corresponding to the i-th row and j-th column sums. For notational convenience (ij) denotes a single numerical datum. The expansion of an entry of Equation (14) becomes:

$$m_{ij} = \mu_{ij} + \sigma_{ij}\left(\alpha_i^R \sum_k \sigma_{ik} + \alpha_j^C \sum_k \sigma_{kj} + \sum_{k\neq j} \sigma_{ik}\alpha_k^R + \sum_{k\neq i} \sigma_{kj}\alpha_k^C\right). \tag{92}$$

Under the substitution $\alpha_i^{R*} = \alpha_i^R \sum_k \sigma_{ik}$ and $\alpha_j^{C*} = \alpha_j^C \sum_k \sigma_{kj}$, the previous expression becomes:

$$m_{ij} = \mu_{ij} + \sigma_{ij}\left(\alpha_i^{R*} + \alpha_j^{C*} + \sum_{k\neq j} \alpha_k^{R*} \frac{\sigma_{ik}}{\sum_l \sigma_{il}} + \sum_{k\neq i} \alpha_k^{C*} \frac{\sigma_{kj}}{\sum_l \sigma_{lj}}\right). \tag{93}$$

If there are many numerical data per accounting identity, it is reasonable to consider that $\sigma_{ik} \ll \sum_l \sigma_{il}$ and that $\sigma_{kj} \ll \sum_l \sigma_{lj}$. Introducing these considerations in the previous expression leads to:

$$m_{ij} = \mu_{ij} + \sigma_{ij}\left(\alpha_i^{R*} + \alpha_i^{C*}\right). \tag{94}$$

The expression above is valid for interior points. The corresponding expressions for row and column sums (labeled respectively with superscripts R and C) are easy to obtain since these data exhibit zero correlations with every other datum. Under the assumption of unitary correlations and a balanced prior, $\sigma_i^R = \sum_k \sigma_{ik}$ and $\sigma_j^C = \sum_k \sigma_{kj}$, the adjustment of row and column sums are:

$$m_i^R = \mu_i^R - \left(\sigma_i^R\right)^2 \alpha_i^R = \mu_i^R - \sigma_i^R \alpha_i^{R*}; \tag{95}$$

$$m_j^C = \mu_j^C - \left(\sigma_j^C\right)^2 \alpha_j^C = \mu_j^C - \sigma_j^C \alpha_j^{C*}. \tag{96}$$

The generalization of the previous expressions to matrix format is given by Equation (17).

References

1. Miller, R.E.; Blair, P.D. *Input-Output Analysis: Foundations and Extensions*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
2. Planting, M.; Guo, J. Increasing the timeliness of US annual input-output accounts. *Econ. Syst. Res.* **2004**, *16*, 157–167.
3. Dalgaard, E.; Gysting, C. An algorithm for balancing commodity-flow systems. *Econ. Syst. Res.* **2004**, *16*, 169–190.
4. Keogh, G.; Quill, P. The construction and analysis of a consistent set of input-output tables for the Irish economy. *J. R. Stat. Soc. Ser. A* **2009**, *172*, 771–788.
5. Lahr, M.L.; de Mesnard, L. Biproportional techniques in input-output analysis: Table updating and structural analysis. *Econ. Syst. Res.* **2004**, *16*, 115–134.
6. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.

7. Deming, E.; Stephan, F.F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **1940**, *11*, 427–444.
8. Lenzen, M.; Gallego, B.; Wood, R. Matrix balancing under conflicting information. *Econ. Syst. Res.* **2009**, *21*, 23–44.
9. Jackson, R.W.; Murray, A.T. Alternative input-output matrix updating formulations. *Econ. Syst. Res.* **2004**, *16*, 135–148.
10. Rampa, G. Using weighted least squares to deflate input-output tables. *Econ. Syst. Res.* **2008**, *20*, 1469–5758.
11. Guilhoto, J.M.; Sesso Filho, U.A. Estimação da matriz insumo-produto a partir de dados preliminares das contas nacionais. *Econ. Apl.* **2005**, *9*, 277–299. (In Portuguese)
12. Guilhoto, J.M.; Sesso Filho, U.A. Estimação da matriz insumo-produto utilizando dados preliminares das contas nacionais: aplicação e análise de indicadores económicos para o Brasil em 2005. *Economia & Tecnologia* **2010**, *6*, 53–62. (In Portuguese)
13. Temurshoev, U.; Webb, C.; Yamano, N. Projection of supply and use tables: Methods and their empirical assessment. *Econ. Syst. Res.* **2011**, *23*, 91–123.
14. Weise, K.; Woger, W. A Bayesian theory of measurement uncertainty. *Meas. Sci. Tech.* **1992**, *4*, 1–11.
15. Laplace, P.S. *Essai Philosophique sur les Probabilités* ; Courcier Imprimeur: Paris, France, 1814. (In French)
16. Jeffreys, H. *Theory of Probability*; Clarendon Press: Oxford, UK, 1939.
17. Jaynes, E.T. *Papers on Probability, Statistics and Statistical Physics*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1983.
18. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
20. Robinson, S.; Cattaneo, A.; El-Said, M. Updating and estimating a social accounting matrix using cross entropy methods. *Econ. Syst. Res.* **2001**, *13*, 45–64.
21. Golan, A.; Judge, G.; Robinson, S. Recovering information from incomplete or partial multisectoral economic data. *Rev. Econ. Stat.* **1994**, *76*, 541–549.
22. Horrace, W.C. On ranking and selection from independent truncated normal distributions. *J. Econ.* **2005**, *126*, 335–354.
23. Sharples, J.J.; Pezzey, J.C.V. Expectations of linear functions with respect to truncated multinormal distributions—with applications to uncertainty analysis in environmental modelling. *Environ. Model. Softw.* **2004**, *16*, 149–156.
24. Raveh, A. On the use of the inverse of the correlation matrix in multivariate data analysis. *Am. Stat.* **1985**, *39*, 39–42.
25. Penrose, R. A generalized inverse for matrices. *Proc. Cambridge Philos. Soc.* **1955**, *51*, 406–413.
26. Golub, G.H.; van Loan, C.F. *Matrix Computations*, 3rd ed.; John Hopkins University Press: Baltimore, MD, USA, 1996.
27. Pereira, X.; Carrascal, A.; Fernandez, M. Impacto económico do turismo receptor através de modelos origem-destino: uma aplicação para a Galiza. In *Modelos Operacionais de Economia*

Regional; Ramos, P., Haddad, E., Eds.; Principia Editora: Parede, Portugal, 2011; pp. 101–121. (In Portuguese)

28. Stone, R.; Champernowne, D.G.; Meade, J.E. The precision of national income estimates. *Rev. Econ. Stud.* **1942**, *9*, 111–125.
29. Stone, R. Multiple classifications in social accounting. *Bulletin de l 'Institut International de Statistique* **1962**, *39*, 215–233.
30. De Mesnard, L. Unicity of biproportion. *SIAM J. Matrix Anal. Appl.* **1994**, *15*, 490–495.
31. Snickars, F.; Weibull, J.W. A minimum information principle theory and practice. *Reg. Sci. Urban Econ.* **1977**, *7*, 137–68.
32. Golan, A.; Vogel, S.J. Estimation of non-stationary social accounting matrix coefficients with supply-side information. *Econ. Syst. Res.* **2000**, *12*, 447–471.
33. Fernandez-Vasquez, E. Recovering matrices of economic flows from incomplete data and a composite Prior. *Entropy* **2010**, *12*, 516–527.
34. Bacharach, M. *Biproportional Matrices and Input-Output Change*; Cambridge University Press: Cambridge, UK, 1970.
35. Byron, R.P. The estimation of large social account matrices. *J. R. Stat. Soc. Ser. A* **1978**, *141*, 359–367.
36. Ploeg, R.V.D. Reliability and the adjustment of sequences of large economic accounts matrices. *J. R. Stat. Soc. Ser. A* **1982**, *145*, 169–186.
37. Berker, T.; Ploeg, F.V.D.; Weale, M. A balanced system of national accounts for the United Kingdom. *Rev. Income Wealth* **1984**, *30*, 461–485.
38. Weale, M. The reconciliation of values, volumes and prices in national accounts. *J. R. Stat. Soc. Ser. A* **1988**, *151*, 211–221.
39. Antonello, P. Simultaneous balancing of input-output tables at current and constant prices with first order vector autocorrelated errors. *Econ. Syst. Res.* **1990**, *2*, 157–172.
40. Lieu, R.; Hicks, R.B.; Bland, C.J. Maximum-entropy in data-analysis with error-carrying constraints. *J. Phys. A* **1987**, *20*, 2379–2388.
41. Lieu, R.; Hicks, R.B. A maximum entropy method when prior information consists of inexact constraints. *Astrophys. J.* **1994**, *422*, 845–849.
42. Lahr, M.L. A strategy for producing hybrid input-output tables. In *Input-Output Analysis: Frontiers and Extensions*; Lahr, M.L., Dietzenbacher, E., Eds.; Palgrave: New York, NY, USA, 2001; pp. 211–242.
43. Gilchrist, D.; Louis, L.S. An algorithm for the consistent inclusion of partial information in the revision of input-output tables. *Econ. Syst. Res.* **2004**, *16*, 149–156.
44. Byron, R.P.; Crossman, P.J.; Hurley, J.E.; Smith, S.C.E. *Balancing Hierarchical Regional Accounting Matrices*; School of Business Discussion Papers, Technical Report 45; Bond University: Gold Coast, Australia, 1993.
45. Chen, B. *A Balanced System of Industry Accounts for the U.S. and Structural Distribution of Statistical Discrepancy*; Working Paper WP2006-8; Bureau of Economic Analysis: Washington, DC, USA, 2006.

46. Rassier, D.; Howells, T.; Morgan, E.; Empey, N.; Roesch, C. *Implementing a Reconciliation and Balancing Model in the U.S. Industry Accounts*; Working Paper WP2007-4; Bureau of Economic Analysis: Washington, DC, USA, 2007.
47. Weidema, B.; Ekvall, T.; Heijung, R. Guidelines for application of deepened and broadened LCA. Available online: http://www.calcasproject.net/ (accessed on 16 January 2014).
48. Bullard, C.W.; Sebald, A.V. Effects of parametric uncertainty and technological change on input-output Models. *Rev. Econ. Stat.* **1977**, *59*, 75–81.
49. Lenzen, M. Errors in conventional and input-output-based life-cycle inventories. *J. Ind. Ecol.* **2001**, *4*, 127–148.
50. Nhambiú, J.O.P. *Avaliação do ciclo de vida de produtos com recurso a quadros económicos de entrada-saída: aplicação a Portugal*; Instituto Superior Técnico, Universidade Técnica de Lisboa: Lisboa, Portugal, 2004. (In Portuguese)
51. Lenzen, M.; Wood, R.; Wiedmann, T. Uncertainty analysis for multi-region input-output models—A case study of the UK's carbon footprint. *Econ. Syst. Res.* **2010**, *22*, 43–63.
52. Yamakawa, A.; Peters, G.P. Using time-series to measure uncertainty in environmental input-output analysis. *Econ. Syst. Res.* **2009**, *21*, 337–362.
53. Díaz, B.; Morillas, A. Incorporating uncertainty in the coefficients and multipliers of an IO table: A case study. *Pap. Reg. Sci.* **2011**, *90*, 845–861.
54. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353.
55. Viertl, R. *Statistical Methods for Fuzzy Data*; Wiley: Chichester, UK, 2011.
56. de Mesnard, L. Negatives in symmetric input-output tables: The impossible quest for the Holy Grail. *Ann. Reg. Sci.* **2011**, *46*, 427–454.
57. Hager, W. Updating the inverse of a matrix. *SIAM Rev.* **1989**, *31*, 221–239.