Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data

João D. F. Rodrigues

To cite this article: João D. F. Rodrigues (2016) Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data, Journal of Business & Economic Statistics, 34:3, 357-367, DOI: 10.1080/07350015.2015.1038545

To link to this article: http://dx.doi.org/10.1080/07350015.2015.1038545

Supplementary materials for this article are available online. Please go to http://tandfonline.com/r/JBES

Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data

João F. D. Rodrigues
Institute of Environmental Sciences (CML), Leiden University, P.O. Box 9518, 2300 RA Leiden, the Netherlands
(*j.rodrigues@cml.leidenuniv.nl*)

Empirical estimates of source statistical economic data such as trade flows, greenhouse gas emissions, or employment figures are always subject to uncertainty (stemming from measurement errors or confidentiality) but information concerning that uncertainty is often missing. This article uses concepts from Bayesian inference and the maximum entropy principle to estimate the prior probability distribution, uncertainty, and correlations of source data when such information is not explicitly provided. In the absence of additional information, an isolated datum is described by a truncated Gaussian distribution, and if an uncertainty estimate is missing, its prior equals the best guess. When the sum of a set of disaggregate data is constrained to match an aggregate datum, it is possible to determine the prior correlations among disaggregate data. If aggregate uncertainty is missing, all prior correlations are positive. If aggregate uncertainty is available, prior correlations can be either all positive, all negative, or a mix of both. An empirical example is presented, which reports relative uncertainties and correlation priors for the County Business Patterns database. In this example, relative uncertainties range from 1% to 80% and 20% of data pairs exhibit correlations below −0.9 or above 0.9. Supplementary materials for this article are available online.

KEY WORDS: Bayesian methods; Maximum entropy principle; Suppressed information.

1. INTRODUCTION

1.1 Motivation

Source statistical economic data are compiled by national statistical offices, and later used in economic analysis and related fields to perform calculations such as changes in employment or carbon emissions embodied in final consumption (Miller and Blair 2009).

Source statistical data are always subject to errors from measurement and processing (Dagum and Cholette 2006; Manski 2014) although only occasionally are such errors reported (Clemen and Winkler 1985; Nicoletti, Peracchi, and Foliano 2011; Cunningham et al. 2012; Meijer, Rohwedder, and Wansbeek 2012). Furthermore, for reasons of statistical confidentiality, very detailed data are sometimes censored (Guldmann 2013). The uncertainty of source statistical data then affects the posterior processing (Stone, Champernowne, and Meade 1942; Ten Raa and Rueda-Cantuche 2003; Wood 2009; Chen 2012) or economic analysis (Hyslop and Imbens 2001; Dietzenbacher 2006), which makes use of that data.

Although information on the stochastic properties of data, that is, their uncertainty and correlation, may not be available, there may exist ancillary information that can be used to obtain estimates of those quantities. For example, it is often the case that statistical data are subject to accounting identities, which express a statistical economic datum as the sum of a set of other data (e.g., employment in a sector equals the sum of employment in every subsector). It may also happen that upper and/or lower bounds can be obtained (e.g., number of jobs is a nonnegative number).

The present article applies the theory of Bayesian inference developed by Jaynes (2003) and the maximum entropy principle (MEP) in particular to determine the stochastic properties of statistical economic data (probability distribution, uncertainty, and correlation) when such information is directly missing but ancillary information is available.

This article presents general formulas that are agnostic concerning either the source of uncertainty or the subsequent use of the generated information. That is, the uncertainty (or imperfect information) can result either from measurement errors or from nondisclosure (or suppressed data), which from the practictioner's point of view are indistinguishable. The formulas derived here can be useful at the stage of data compilation if, for example, the resulting priors are used to improve the balancing conflicting estimates; or they can be useful at the later stage of studying uncertainty propagation.

1.2 Problem Formulation

According to Weise and Woger (1992), a numerical datum subject to measurement error is described by a random variable t and a probability distribution $p(q)$, which expresses the belief that the "true" value of the poorly known datum t takes realization q.

Besides the probability distribution $p(q)$, the datum is characterized by a *best guess* or expectation, $m = \mathrm{E}[t]$, and by an *uncertainty* estimate or standard deviation, $s = \sqrt{\mathrm{var}[t]}$. The best guess is the observable quantity, for example, the published point estimate. The uncertainty expresses a degree of confidence

Journal of Business & Economic Statistics
July 2016, Vol. 34, No. 3
DOI: 10.1080/07350015.2015.1038545

that the best guess is close to the true value of the statistical economic datum.

Furthermore, statistical economic data are often constrained by *accounting identities* (e.g., total output equals the sum of sales to different institutional sectors), whose general formulation is

$$t_0 = \sum_{i=1}^{n} t_i, \tag{1.1}$$

where t_0 is an *aggregate* datum and t_i where $i > 0$ are *disaggregate* data (Hendry and Hubrich 2011). The existence of accounting identities implies that there are *correlations*, r_{ij}, which show how pairs of data co-vary:

$$r_{ij} = \frac{\text{cov}[t_i, t_j]}{\sqrt{\text{var}[t_i]}\sqrt{\text{var}[t_i]}}.$$

If a correlation is zero, $r_{ij} = 0$, then the data points are uncorrelated, meaning that the realization that datum i takes does not affect the realization of j. If a correlation is one, $r_{ij} = 1$, then the data points are perfectly correlated, and knowing the realization of i determines the realization of j. For example, if both i and j have Gaussian distribution, perfect correlation implies that if $q_i = m_i + zs_i$, then $q_j = m_j + zs_j$, where z is a real number. In general, a correlation can take any value in the range $(-1, 1)$.

This article uses the MEP (Jaynes 1957) to determine the prior probability distribution, uncertainty, and correlation of statistical data. A prior parameter is a parameter for which no previous estimate was available but which can be obtained by inductive inference from contextual information (Jaynes 2003).

In the present work, this problem is addressed in the context of statistical economic data and the following concrete questions are addressed:

1. What is the prior probability distribution, $p(q)$, which characterizes a datum in isolation, t, when only a best guess, m, and uncertainty estimate, s, are available?
2. What is the prior uncertainty estimate, s, which characterizes a datum in isolation, t, when only a best guess, m, is available?
3. What is the prior correlation, r_{ij}, which characterizes two data points, t_i and t_j, constrained by an accounting identity, Equation (1.1)?
4. What is the prior uncertainty, s_0 of an aggregate datum, t_0, when only the uncertainties, s_i, of disaggregate data points, t_i with $i > 0$, are available?

1.3 Outline of the Article

Sections 2 and 3 present the theoretical and computational developments, which are organized as follows.

Section 2 determines the probability distribution that characterizes an isolated datum, answering the first two questions presented in Section 1.2.

Section 3 determines the probability distribution and the correlation among data connected by an accounting identity, and thus addresses the other questions presented in Section 1.2.

To clarify the results the theoretical sections are accompanied by illustrative examples. Section 2 describes the probability density of a strictly positive datum with unitary best guess, as a function of relative uncertainty. Section 3 studies the different correlation patterns that emerge in an accounting identity with only three disaggregate data points.

Section 4 presents a real-world application that shows the range of uncertainties and correlations displayed by a statistical economic dataset.

Section 5 concludes. Auxiliary material (Appendices A–F) is reported as supplementary information (available online).

2. UNCERTAINTY OF AN ISOLATED DATUM

2.1 Review and Assumptions

In the past, different families of probability distributions have been assigned to statistical data. For example, Golan, Judge, and Robinson (1994) considered a discrete uniform distribution, Golan and Vogel (2000) considered a discrete triangular distribution, Dietzenbacher (2006) considered a gamma distribution, Díaz and Morillas (2011) considered a beta distribution, and Lenzen, Wood, and Wiedmann (2010, p. 46) considered a log-normal distribution. Nonetheless, the most popular probability distribution used is the nontruncated symmetric Gaussian (Lenzen, Wood, and Wiedmann 2010, p. 44).

In contrast to this literature, in the present article a probability distribution is not postulated but is instead derived from first principles. According to the Bayesian paradigm (Jaynes 2003), the best inference takes into account all available information and no other. This implies that the prior probability distribution of a statistical economic datum is obtained by the MEP. This principle, formulated by Jaynes (1957) and based on the work of Shannon (1948), states that the least informative probability distribution consistent with a given set of constraints is the one which maximizes entropy. Thus, if an unknown datum can take discrete values q_j, with $j = 1, \ldots, n_L$, and its first n_M moments, $M_i = \sum_{j=1}^{n_L} (q_j)^i p(q_j)$, are known, then its least informative probability distribution, $p(q_j)$, maximizes the Lagrangian

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

The first term on the right-hand side is the entropy of distribution $p(q_j)$, and the second term is the set of constraints, where the λ_i's are Lagrange multipliers. The MEP determines a prior, when it is possible to express the available information in terms of moments, by making sure that no other information is being used (i.e., maximizing ignorance or uncertainty).

Statistical economic data (monetary transactions, employment, carbon emissions, etc.) are reported by statistical offices as real numbers with a finite number of digits (e.g., multiples of 10^3 euros, full-time jobs, or tons of CO_2). Furthermore, the precision with which the data are reported is usually independent of the scale. For example, with a precision of two decimal cases the data points 1.2345 and 123.456789 are reported as 1.23 and 123.46 (due to roundoff).

Under these conditions, it is reasonable to approximate $p(q)$ by a continuous distribution and to replace the discrete version of

the MEP, described above, by differential entropy with uniform measure. If the precision is not uniform but exponential, for example, if the possible values for the numerical datum are 1, 2, 4, 8, etc., then a different measure should be used. This choice would lead to a probability density function different from the one derived below (Frank and Smith 2010).

In this work it is assumed that the realization, q, of a statistical economic datum, t, can take any value in the range $(0, \infty)$ as most statistical economic data are nonnegative by definition (e.g., an economic transaction or GHG emissions). A negative number may appear for conventional reasons (e.g., a positive number as an input and a negative number as an output), and can therefore be converted to a positive number by means of a topological transformation, as described in appendix A.1 of Rodrigues (2014).

2.2 Analytical Solutions

Following Weise and Woger (1992), the *best guess*, m, and *uncertainty*, s, of the source data are interpreted as the expected value, $\mathrm{E}(t) = m$, and the standard deviation, $\mathrm{var}(t) = s^2$, where $\mathrm{E}(f(t)) = \int_0^\infty dq p(q) f(q)$ and $\mathrm{var}(t) = \mathrm{E}(t^2) - \mathrm{E}(t)^2$. Under these conditions the Lagrangian is

$$L = -\int_0^\infty dq p(q) \ln(p(q)) + \lambda(\mathrm{E}(1) - 1) \qquad (2.1)$$
$$+\alpha(\mathrm{E}(t) - m) + \beta\left(\mathrm{E}(t^2) - \mathrm{E}(t)^2 - s^2\right).$$

The first term on the right-hand side of Equation (2.1) is the differential entropy of the unknown distribution. The remaining terms on the right-hand side of Equation (2.1) are the set of known constraints: the *zeroth-order* constraint is the normalization, the *first-order* constraint is the expected value, and the *second-order* constraint is the variance. λ, α, and β are the respective Lagrange multipliers. If the uncertainty is not known, then the second-order constraint is removed from Equation (2.1), and $\beta = 0$.

According to Dowson and Wragg (1973), the maximization of Equation (2.1) with respect to $p(q)$ leads to

$$p(q) = C \exp\left(\alpha q + \beta\left(q^2 - 2mq\right)\right), \qquad (2.2)$$

where C is a constant. Since Equation (2.1) defines a concave function, differentiation yields a unique maximum. There are two cases, depending on whether an uncertainty estimate is available or not.

If uncertainty is not known, $\beta = 0$, and Equation (2.2) leads to an exponential distribution

$$p(q) = \alpha e^{-\alpha q}. \qquad (2.3)$$

The expected value and the standard deviation of the exponential distribution are $m = s = 1/\alpha$, so if no uncertainty estimate is provided, the MEP determines a prior uncertainty $s = m$.

If an uncertainty estimate is available, Equation (2.2) leads to a truncated Gaussian distribution

$$p(q) = \frac{1}{Z}\frac{1}{\sqrt{2\pi\tilde{s}^2}} \exp\left(-\frac{(q-\tilde{m})^2}{2\tilde{s}^2}\right), \qquad (2.4)$$

with the substitution $2\beta = 1/\tilde{s}^2$ and $\alpha - 2m\beta = -\tilde{m}/\tilde{s}^2$, where Z is a normalization constant. Note that since this distribution is truncated, the *Gaussian* parameters $\tilde{m}$ and $\tilde{s}^2$ are *not* the *observable* expectation and variance of the distribution, m and s^2. The properties of the truncated Gaussian distribution have been studied in the past (Cohen 1950; Castillo 1994) but unfortunately, there is no closed form analytical expression connecting (m, s) and $(\tilde{m}, \tilde{s})$ (Tallis 1961). Using the inverse Mills ratio it is possible to express observables as a function of Gaussian parameters (Greene 2008), but the reverse is not true. Johnson and Kotz (1970, pp. 81–87) reviewed several methods to perform this conversion, including the method of Pearson and Lee (1908), but all of these methods involved numerical root finding. Appendix A (see online supplementary materials) presents expressions that allow for the explicit conversion from parameters to observables and vice versa.

2.3 Transition Between Solutions

There is a smooth transition between the first- and second-order MEP distributions (Cover and Thomas 1991; Castillo 1994). If the relative uncertainty, s/m, is small, the truncated Gaussian distribution is well approximated by its nontruncated cognate. However, as relative uncertainty increases, the probability mass gets increasingly skewed to the left, until it becomes indistinguishable from the exponential distribution, when $s/m \simeq 1$ (Dowson and Wragg 1973).

Figure 1 shows the probability density function of the truncated Gaussian distribution, for different levels of observable relative uncertainty. When relative uncertainty is below 0.3, the truncated Gaussian is well-approximated by its nontruncated cognate. When relative uncertainty rises to 0.75, the peak of the function smashes against the zero boundary and the function becomes monotonic.

The limit behavior of high relative uncertainty can be deduced analytically. Let the probability density of the truncated Gaussian (Equation (2.4)) be expanded as

$$p(q) = C \exp\left(-\frac{(q-\tilde{m})^2}{2\tilde{s}^2}\right) = C \exp\left(-\frac{q^2}{2\tilde{s}^2} + \frac{2q\tilde{m}}{2\tilde{s}^2} - \frac{\tilde{m}^2}{2\tilde{s}^2}\right),$$

where the C's in the previous and following expression are appropriately chosen constants. In the limit case of high uncertainty, $\tilde{m} \to -\infty$ and $\tilde{s} \to \infty$, but the bulk of probability mass is constrained in the lower positive range, $0 < q \ll \infty$. Under these conditions, $q^2/\tilde{s}^2 \simeq 0$ and $\tilde{m}^2/\tilde{s}^2$ is a constant, so

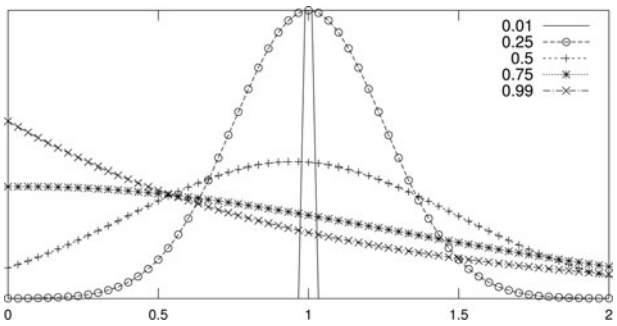


Figure 1. Probability density function of the truncated Gaussian distribution, for five different levels of observable relative uncertainty, s/m.

the previous expression simplifies to

$$\pi(q) = C \exp\left(-\frac{|\tilde{m}|}{\tilde{s}^2} q\right).$$

That is, the far-right tail of a truncated Gaussian distribution exhibits an exponential shape and, in this limit case, there is an explicit link between Gaussian and observable parameters: $|\tilde{m}|/\tilde{s}^2 = 1/m = 1/s$.

Thus, the prior relative uncertainty of an isolated numerical datum, $u = s/m$ is bound, $0 \leq u \leq 1$, and if no uncertainty estimate is provided, then prior relative uncertainty is unitary, $u = 1$ and $s = m$.

3. CORRELATIONS AMONG CONNECTED DATA

3.1 Constraints on Aggregate Uncertainty

Thus far, this article studied the properties of a statistical economic datum in isolation. However, statistical economic data are often connected to one another through accounting identities, Equation (1.1), linking one aggregate datum to several disaggregate data.

From standard probability theory, it follows from Equation (1.1) that

$$m_0 = \sum_{i=1}^{n} m_i; \tag{3.1}$$

$$s_0^2 = \sum_{i=1}^{n} s_i^2 + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} r_{ij} s_i s_j, \tag{3.2}$$

where r_{ij} in Equation (3.2) is the correlation between disaggregate data i and j. Thus, Equation (3.2) places constraints on the correlation between disaggregate data and the uncertainty of the aggregate datum.

Recall that correlations have the following properties: $r_{ii} = 1$, $r_{ij} = r_{ji}$, and $-1 \leq r_{ij} \leq 1$. The presence of correlations defines an uncertainty range for the uncertainty of the aggregate datum, which is narrower than the uncertainty range of disaggregate data.

Consider that all correlations have the highest possible value, $r_{ij} = 1$. Equation (3.2) becomes

$$s_0^2 = \sum_{i=1}^{n} s_i^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} s_i s_j = \left(\sum_{i=1}^{n} s_i\right)^2.$$

Because correlations cannot be larger than one, the previous expression implies that an upper bound for aggregate uncertainty is

$$s_{\max} = \sum_{i=1}^{n} s_i. \tag{3.3}$$

Combining Equations (3.3) and (3.1) leads to the observation that the upper bound of the relative uncertainty of the aggregate datum, $u_0 = s_0/m_0$, is the average of the relative uncertainties of the disaggregate data, $u_i = s_i/m_i$:

$$u_0 \leq \sum_{i=1}^{n} \frac{m_i}{m_0} u_i \leq 1.$$

The previous expression defines an upper limit for the uncertainty of an aggregate datum, which may be lower than the upper uncertainty limit of a disaggregate datum, given in Section 2 as $u_i \leq 1$. Heijungs and Suh (2002, p. 140–144) have previously identified this constraint in the field of life-cycle assessment.

Likewise, there is a lower bound for aggregate uncertainty:

$$s_{\min} = \max\left\{0, s_1 - \sum_{i=2}^{n} s_i\right\}, \tag{3.4}$$

where it is assumed that disaggregate uncertainties are ordered by decreasing size, so that s_1 is the largest uncertainty.

This lower bound arises because the lowest possible aggregate uncertainty of two disaggregate data occurs when the correlation between them is -1. Furthermore, according Equation (3.3), the highest uncertainty of the subset of $i = 2, \ldots, n$ disaggregate data is obtained when they are all perfectly correlated.

Hence, if the uncertainty of the largest disaggregate datum is larger than the sum of the uncertainties of all other disaggregate data, there is a positive lower bound for aggregate uncertainty.

Finally, a situation of particular interest is the aggregate correlation that occurs when all correlations are zero

$$s_{\text{zero}} = \sqrt{\textstyle\sum_{i=1}^{n} s_i^2}. \tag{3.5}$$

Values $s_{\min}$ and $s_{\max}$ are the lower and upper bound for aggregate uncertainty, s_0, and the configuration of prior correlations will depend on how s_0 is positioned in relation to s_{zero}.

3.2 Determination of Correlations

According to the Bayesian paradigm, the best inference takes into account all available information and no other. Accounting identities are a very strong piece of information to which the assignment of priors must conform. In fact, the combination of accounting identities and the MEP allows the determination of correlation priors.

Appendix B (see online supplementary materials) presents the derivation of the analytical solution of correlation priors constrained by Equation (1.1) when all uncertainties are known. The solution is

$$\frac{\tilde{w}_{ij}}{\tilde{s}_i \tilde{s}_j} = \beta, \tag{3.6}$$

for every $i \neq j$, where β is the Lagrange parameter and $\tilde{w}_{ij}$ is the (i, j) entry of the inverse correlation matrix $\tilde{\mathbf{W}} = \tilde{\mathbf{S}}^{-1}$. Recall that all parameters adjoined by a tilde,˜, are Gaussian parameters, which differ from observable parameters when relative uncertainty is high.

Appendix B (see online supplementary materials) also addresses the problem of determining correlations when the aggregate uncertainty prior is unknown. In this case, correlation priors are given by Equation (3.6) and the aggregate datum prior uncertainty is given by

$$\frac{1}{\tilde{s}_0^2} = -\beta. \tag{3.7}$$

Notice that, in comparison to Equation (3.6), the right-hand side of Equation (3.7) has a minus sign.

Let us consider that all uncertainties are known and that all correlations are unknown. Furthermore, consider that relative

uncertainties are low, so that observable and Gaussian parameters are interchangeable. Substitution of Equation (3.6) in the matrix product $\mathbf{I} = \mathbf{WR}$ leads to the following constraints:

$$1 = w_{ii}s_i^2 + \beta \sum_{k \neq i} r_{ik}s_i s_k;$$

$$0 = w_{ii}r_{ij}s_i s_j + \beta s_j^2 + \beta \sum_{k \neq i,j} r_{kj}s_k s_j.$$

In the previous and following expressions, summation is always in the range $k = 1, \ldots, n$, except for the referred iterator (e.g., $k \neq i$). Using the first expression to eliminate w_{ii} from the second expression leads to

$$0 = r_{ij} + \beta \left(s_i s_j + s_i \sum_{k \neq i,j} r_{kj}s_k - r_{ij}s_i \sum_{k \neq i} r_{ik}s_k \right). \quad (3.8)$$

The full algorithm for the determination of prior correlations is a two-stage Newton method (Press et al. 2007). First, the root of Equation (3.8) can be solved for a given parameter β. Then the parameter β itself is obtained by finding the root of Equation (3.2).

3.3 Prior Correlations and Missing Aggregate Uncertainty

Few studies consider nonzero prior correlations among statistical economic data in a given year, such as Weale (1988) or Antonello (1990) in the case of data balancing, Rypdal and Zhang (2000) or Flugsrud and Hoem (2011) in the case of greenhouse gas emissions, or Ballantyne et al. (2012) in the case of time-series carbon concentrations. Yet, most work on the uncertainty of calculations based on statistical economic data considers only zero correlations (Lenzen 2001; Dietzenbacher 2006; Rampa 2008).

On this subject, Rassier et al. (2007, p. 9) stated that "given the lack of information regarding correlations among the initial estimates, covariance measures are assumed to be zero. While this assumption results in an estimator that is less than efficient, the inefficiency is less than may be introduced if the correlations are incorrectly determined." Also, according to Dagum and Cholette (2006), one justification of zero correlation is that for a large system the covariance matrix may become ill-conditioned.

This situation is very different for the case of time-series data, in which a large literature for the estimation of correlations and their subsequent use in statistical analysis is available (Chow and Lin 1971a, 1971b; Cholette and Dagum 1994; Dagum, Cholette, and Chen 1998; Engle and Kelly 2012).

Consider that there are independent estimates of each disaggregate best guess, m_i, that there may or may not be independent estimates of disaggregate uncertainties, s_i (it will not affect the remainder of the analysis), and that no estimate of the aggregate uncertainty, s_0, is available. The value of prior correlations, r_{ij}, will depend on whether an independent estimate of the aggregate best guess, m_0, is available or not. The formal analysis of this matter is reported in Appendix C (see online supplementary materials). An informal discussion of the results is now presented.

If no independent prior for m_0 is available, then the correlation data is obtained by maximizing the entropy of joint disaggregate data *but not* of the aggregate datum: it must be so because the latter is, literally, outside the scope of the study. In this case, the correlations are zero, $r_{ij} = 0$, and the aggregate relative uncertainty is obtained from Equation (3.5).

If an independent prior for m_0 is available, then the correlation data is obtained by maximizing the entropy of joint disaggregate data *and* of the aggregate datum. Whereas the former is maximized when correlations are zero, the latter is maximized when correlations are unitary. Because both are monotonic in the range $0 < r_{ij} < 1$, it follows that the MEP solution occurs when all correlations are positive.

Hence, the conventional assumption of zero correlations corresponds to a particular empirical situation of interest (absence of an initial estimate of the aggregate best guess), and is consistent with the constraint posed by the accounting identity (Equation (3.2)) if aggregate uncertainty conforms to Equation (3.5).

However, in the conventional literature the information content of accounting identities is sometimes overlooked. For example, both Golan, Judge, and Robinson (1994), in their generalized cross-entropy problem, and Lenzen (2011, p. 76) considered that disaggregate data have positive uncertainty, that aggregate data have zero uncertainty, and that all correlations among disaggregate data are zero.

These assumptions are mutually inconsistent, since they violate Equation (3.2). On the one hand, if correlations are all zero, then the uncertainty of the aggregate datum is not independent but should be obtained by Equation (3.5). On the other hand, if the uncertainty of the aggregate datum is zero, it follows from the analysis above that the correlations between disaggregate data must be negative on average. Furthermore, Section 3.1 found that depending on the configuration of disaggregate uncertainties, it may even be impossible for the aggregate uncertainty to be zero.

3.4 Qualitative Patterns

The precise configuration of prior correlations will naturally depend on the uncertainty values, but several qualitative patterns hold in general, defined by the dispersion in disaggregate uncertainties and, especially, by the gap between the largest disaggregate datum and the sum of all other disaggregate data.

To illustrate these qualitative patterns we now present a series of examples. Each example consists of an accounting identity with three disaggregate data and is defined by a particular combination of disaggregate uncertainties, s_1, s_2, and s_3. For each example the configuration of correlations is shown in a figure, as a function of aggregate relative uncertainty.

To aid interpretation, key values of aggregate uncertainty are reported: $s_{\max}$, given by Equation (3.3), is the upper bound of aggregate uncertainty; $s_{\min}$, given by Equation (3.4), is the lower bound of aggregate uncertainty; s_{zero}, given by Equation (3.5), is the zero-correlation uncertainty; and s_{mep}, given by Equation (3.7), is the aggregate uncertainty prior when an aggregate best guess is initially available.

Figure 2 illustrates what happens when the largest disaggregate uncertainty exceeds the sum of the remainder disaggregate uncertainty. In this case, there is a nonzero lower bound for

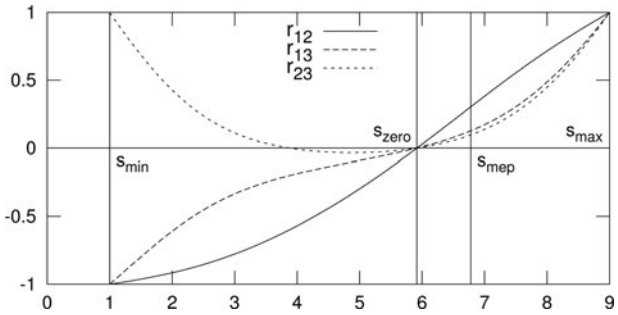


Figure 2. Correlations between disaggregate data as a function of aggregate uncertainty, when disaggregate uncertainties are $s_1 = 5, s_2 = 3$, and $s_3 = 1$.

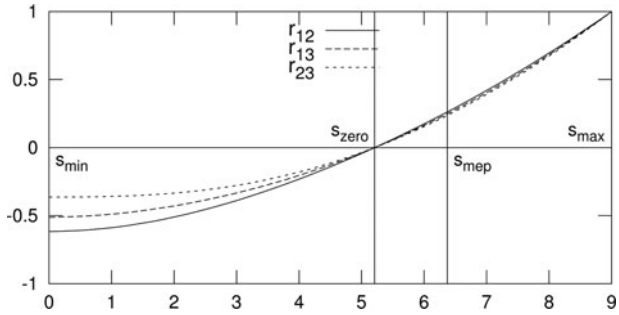


Figure 4. Correlations between disaggregate data as a function of aggregate uncertainty, when disaggregate uncertainties are $s_1 = 3.25$, $s_2 = 3$, and $s_3 = 2.75$.

aggregate uncertainty, $s_{\min} > 0$, such that when aggregate uncertainty falls far below the zero-correlation uncertainty (at 4, in this particular case) there is a divergence among disaggregate correlations. Correlations between the lower uncertainty and the largest uncertainty become negative, while the correlation among lower uncertainty data become positive. When aggregate uncertainty takes the minimum value, these correlations become, respectively, -1 and 1.

Figure 3 shows the results when there is still some variation between disaggregate uncertainties but the largest disaggregate uncertainty does not exceed the sum of the remainder. Here the situation is similar to the previous example except in the limit of low aggregate uncertainty, in which the correlations between disaggregate data do not become perfectly correlated or perfectly anti-correlated.

Figure 4 shows the results when disaggregate uncertainties are very similar. In this case the pattern is more regular, with all correlations remaining negative if aggregate uncertainty is below the zero-correlation uncertainty.

Appendix D (see online supplementary materials) presents a more exhaustive characterization of the different correlation patterns that emerge from the combination of disaggregate uncertainties.

4. EMPIRICAL ILLUSTRATION

4.1 Scope

The goal of this section is to present an empirical illustration of the correlation patterns that are obtained using the MEP. These patterns will be contrasted with the conventional assumption of zero correlations.

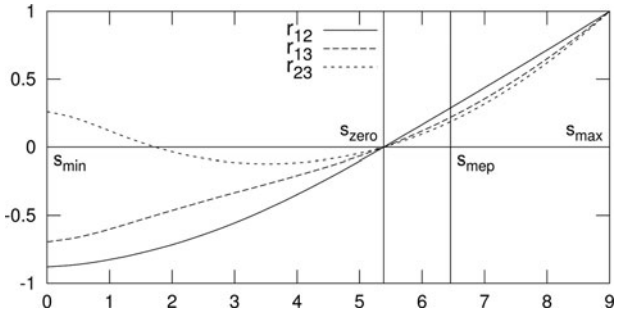


Figure 3. Correlations between disaggregate data as a function of aggregate uncertainty, when disaggregate uncertainties are $s_1 = 4, s_2 = 3$, and $s_3 = 2$.

It may happen that, for a particular accounting identity, the uncertainties satisfy Equation (3.5), in which case MEP correlations are zero. If that condition is not satisfied, then the conventional assumption does not hold and nonzero correlations must be explicitly taken into account.

The example will also look at relative uncertainties. If relative uncertainties is less than one-third, the MEP probability distribution (the truncated Gaussian) is indistinguishable from the conventional nontruncated Gaussian. If relative uncertainty is higher than one-third, then the nontruncated Gaussian is no longer acceptable, as it would violate the nonnegativity condition.

The example uses employment data of the County Business Patterns (CBP) for the Autauga county of the state of Alabama in the year 2000. This dataset was chosen for several reasons: it is of open access and thus allows validation of the present results by a third party; it is amenable to a transparent processing procedure; it is topologically simple (one employment estimate per industry, county and year) and can thus be described briefly; and it offers a wide variety of empirical patterns (accounting identities with different number of data points and uncertainties).

This type of regional employment data is used for different purposes, including the calibration of regional economic models (Treyz and Stevens 1985; Lahr and Stevens 2002).

4.2 Data Source and Processing

The CBP database (*http://www.census.gov/econ/cbp/*), maintained by the U.S. Census Bureau, reports the number of employees per industry following the NAICS 2002 classification scheme (*http://www.census.gov/eos/www/naics/*) up to six digits. These data are reported for each county of every state in the United States.

The remainder of this subsection describes a particular processing procedure for the extraction of best guess and uncertainty estimates from this source data, which is not only conceptually sound but also simple and transparent enough to allow for easy reproducibility. More sophisticated methods can be found in Fischetti and Salazar (2005), Isserman and Westervelt (2006), Zhang and Guldmann (2009), Chen (2012), Guldmann (2013), and Zhang and Guldmann (2015).

The CBP database reports a single employment figure (the number of jobs in mid-March) for industries with a large number of establishments. This number is taken as the prior best guess. For such industries, two wage values are reported: first quarter payroll, FQP_i, and annual payroll, AP_i. Payroll is used as a proxy for the number of jobs so that the relative uncertainty of the number of jobs is $|4 \times \mathrm{FQP}_i - \mathrm{AP}_i|/|4 \times \mathrm{FQP}_i + \mathrm{AP}_i|$.

To protect confidentiality (Doyle et al. 2001), the employment data of industries with a small number of establishments is flagged and the total number of employees is not disclosed but instead a range is presented (1–19, 20–99, etc.). Furthermore, for each industry (whether flagged or not), the dataset also indicates the number of establishments by employee size class (1–4, 5–9, 10–19, etc.).

For flagged industries, a lower and an upper bound, LB_i and UB_i, were obtained as the narrower bound defined by the industry flag and the employee size classes. As an example, consider industry 1133 (logging) of the Autauga county of the state of Alabama in the year 2000. According to the industry flag, there are between 0 and 19 employees in this industry. However, the industry contains three establishments, each with a number of employees in the range 1 to 4. Hence, for this industry, $\mathrm{LB}_i = 3$ and $\mathrm{UB}_i = 12$. From the lower and upper bound, the best guess was obtained as $(\mathrm{LB}_i + \mathrm{UB}_i)/2$ and relative uncertainty as $|\mathrm{LB}_i - \mathrm{UB}_i|/(\mathrm{LB}_i + \mathrm{UB}_i)$.

In a few instances, the source information between the industry flag and the employee size to class was found to be inconsistent, that is, they expressed disjoint sets. This probably resulted from a misspecification problem (Abowd and Vilhuber 2005), and in this case the inconsistency was solved by manually adjusting one of the flags.

The NAICS hierarchy provides a set of accounting identities that constrain employment values between industries of sequential digit levels. For example, with the priors obtained using the procedure above, industry 113 (forestry and logging) employed 11 ± 0.25 workers, which were divided into 2.5 ± 2 jobs in industry 1131 (timber tract operations) and 7.5 ± 4.5 jobs in industry 1133 (logging).

The set of prior best guesses thus obtained was found to be inconsistent (i.e., first-moment accounting identities did not hold) and was balanced using the linear method of Rodrigues (2014). The linear method is able to handle constraints of arbitrary structure, a hierarchy of information quality, and reliability weights. The method is iterative and, when reliability weights are identical for all elements and constraints are row and column sums, reduces to conventional biproportional adjustment (Lahr and Mesnard 2004). The method yields a set of balanced posterior best guesses and preserves relative uncertainties. Technical details are summarized in Appendix E (see online supplementary materials). Ideally, information on correlations would even be used in the balancing procedure itself, as in the generalized least squares method of Rodrigues (2014). However, that would make the balancing algorithm computationally and conceptually more complex.

The uncertainties of aggregate data were then adjusted to conform with the upper and lower bounds described in Section 3.1. The resulting set of valid uncertainties and balanced best guesses was used as input data to estimate the maximum entropy prior correlations.

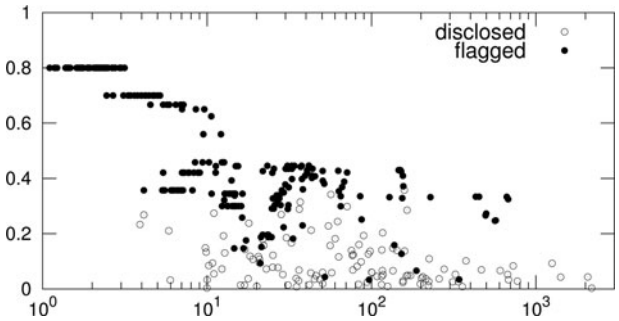


Figure 5. Relative uncertainty of balanced disclosed and flagged employment estimates as a function of the best guess of the number of employees.

To test the robustness of the empirical patterns to the processing procedure, several variations on the definition of best guess priors and the balancing algorithm were considered, as described in Appendix F (see online supplementary materials).

4.3 General Results

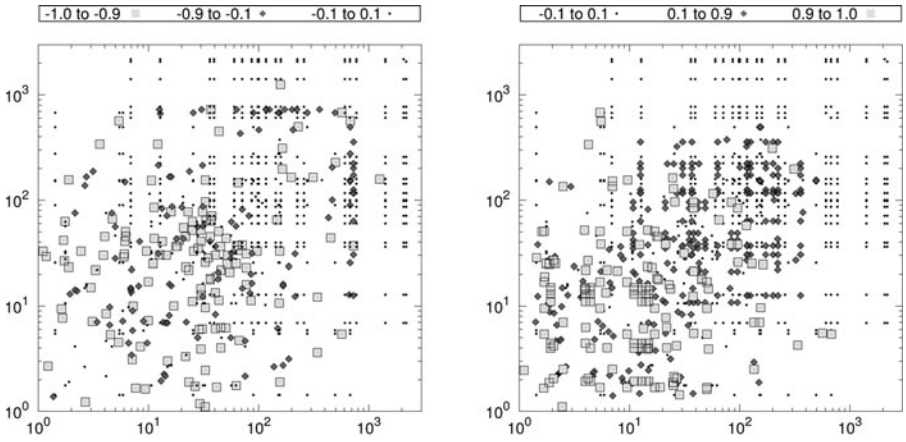
The dataset under study reports employment estimates for different industries scattered across six NAICS digit levels. The county contains 20 two-digit, 75 three-digit, 171 four-digit, 245 five-digit, and 257 six-digit industries. There is a total of 388 nonredundant industries (since often a higher-level industry branches into a single lower-level industry), of which 118 have disclosed employment data and 270 are flagged. Figure 5 shows the relative uncertainty of balanced disclosed and flagged industry employment estimates, as a function of the best guess of the number of employees.

In the balanced configuration, the best guesses of employment figures of disclosed industries are scattered from 4 to around 2000 employees (not counting the county total), with 5% of industries with more than a thousand employees, 50% having more than 70 and 80% having more than 20. The corresponding values for flagged data are scattered from 1 to around 600 employees, with 50% of industries having a best guess smaller than eight and 90% smaller than 100.

Sixty percent of disclosed industries have relative uncertainty below 0.1, while 90% of disclosed industries have relative uncertainty below 0.2. The lowest and highest uncertainties of a disclosed industry are 0.0035 and 0.36, respectively. Forty percent of flagged industries have an uncertainty higher than 0.6, and 90% of flagged industries have an uncertainty higher than 0.3. The lowest and highest uncertainties of a flagged industry are 0.03 and 0.8, respectively.

All disclosed employment data have a relative uncertainty below (or close to) one-third. Hence, according to Section 2.3 they are well described by the Gaussian approximation. Most flagged data is not well approximated by the Gaussian distribution, although they are also not well approximated by the exponential limit either (uncertainty above 90%), falling somewhere in between.

There is a total of 131 nonredundant accounting identities, connecting a higher-digit parent industry to more than one



(a) Negative correlations. (b) Positive correlations.

Figure 6. Correlations by range, as a function of the best guess estimates of the number of employees.

lower-digit daughter industries, of which 77 include two disaggregate data points, 34 include three, 17 include between four and eight, and the largest three include 12, 15, and 20. Because the number of distinct correlations constrained by an accounting identity is $n(n-1)/2$, where n is the number of disaggregate data points, the set of accounting identities defines a total of 722 (possibly nonzero) correlations between industry employment figures in the Autauga county of Alabama in the year 2000.

Figure 6(a) and 6(b) show the correlations between different data points, as a function of best guesses. Eight percent of correlations are lower than −0.9, 11% are between −0.9 and −0.1, 47% are between −0.1 and 0.1, 22% between 0.1 and 0.9, and 12% higher than 0.9. Hence, although almost half of all correlations are close to zero, there is a significant number that is quite different from zero.

In Figure 6(a) and 6(b), it is possible to see that the correlations that are closer to zero are scattered all over the range of best guess values. However, correlations that are significantly different from zero occur mostly between industries with small employment best guesses. Industries whose correlation is lower than −0.9 have a median employment best guess of 31 jobs, those whose correlations are between −0.9 and −0.1 of 36, between −0.1 and 0.1 of 86, between 0.1 and 0.9 of 35, and higher than 0.9 of 9. Hence, very high correlations are concentrated among industries with a small number of employees.

Table 1. Properties of selected aggregate industries

Code	u_0	m_0	s_0	$s_{\min}$	s_{zero}	s_{mep}	$s_{\max}$
621	0.03	311.74	8.18	0.00	5.90	7.43	11.70
48	0.10	86.51	8.49	5.31	11.18	12.10	15.78
4441	0.01	115.13	0.83	0.00	22.21	25.90	36.47
4461	0.04	201.30	7.83	0.00	10.82	13.11	19.88

NOTE: Code = NAICS codes (see Table 6); u_0 = relative uncertainty; m_0 = best guess; s_0 = absolute uncertainty; $s_{\min}$ = lower bound; s_{zero} = zero-correlation uncertainty; s_{mep} = maximum-entropy uncertainty; $s_{\max}$ = upper bound.

Table 2. Properties of disaggregate industries constrained by NAICS industry 621 (Ambulatory health care services)

Code	u	m	s	R	6213	6214	6211	6212
6213	0.16	22.07	3.48	6213	1.00	0.35	0.33	0.31
6214	0.12	25.20	3.07	6214	0.35	1.00	0.31	0.29
6211	0.01	221.73	2.66	6211	0.33	0.31	1.00	0.27
6212	0.06	42.75	2.48	6212	0.31	0.29	0.27	1.00

NOTE: Code = NAICS codes (see Table 6); u = relative uncertainty; m = best guesses; s = absolute uncertainty; = absolute uncertainties; R = correlations.

Table 3. Properties of disaggregate industries constrained by NAICS industry 48 (Transportation and warehousing)

Code	u	m	s	R	484	485	481	488
484	0.14	75.56	10.55	484	1.00	−0.59	−0.36	−0.06
485	0.70	4.67	3.27	485	−0.59	1.00	0.13	0.02
481	0.80	2.16	1.73	481	−0.36	0.13	1.00	0.01
488	0.06	4.13	0.24	488	−0.06	0.02	0.01	1.00

NOTE: Same description as in Table 2.

Table 4. Properties of disaggregate industries constrained by NAICS industry 4441 (building material and supplies dealers)

Code	u	m	s	R	44413	44411	44412	44419
44413	0.43	40.56	17.44	44413	1.00	−0.98	−0.77	−0.68
44411	0.43	30.05	13.06	44411	−0.98	1.00	0.66	0.58
44412	0.42	8.31	3.50	44412	−0.77	0.66	1.00	0.46
44419	0.07	36.21	2.47	44419	−0.68	0.58	0.46	1.00

NOTE: Same description as in Table 2.

Table 5. Properties of disaggregate industries constrained by NAICS industry 4461 (health and personal care stores)

Code	u	m	s	R	44611	44619	44612	44613
44611	0.05	166.92	8.20	44611	1.00	−0.30	−0.28	−0.12
44619	0.30	16.12	4.84	44619	−0.30	1.00	−0.10	−0.04
44612	0.30	15.52	4.66	44612	−0.28	−0.10	1.00	−0.04
44613	0.80	2.74	2.19	44613	−0.12	−0.04	−0.04	1.00

NOTE: Same description as in Table 2.

Table 6. Description of the NAICS codes of selected accounting identities

Type	Code	Description
Aggregate	621	Ambulatory health care services
Disaggregate	6211	Offices of other health practitioners
	6212	Outpatient care centers
	6213	Offices of physicians (exc mental health)
	6214	Offices of dentists
Aggregate	48	Transportation and warehousing
Disaggregate	481	Transit and ground passenger transportation
	484	Other nonscheduled air transportation
	485	Truck transportation
	488	Transportation support activities
Aggregate	4441	Building material and supplies dealers
Disaggregate	44411	Paint and wallpaper stores
	44412	Hardware stores
	44413	Home centers
	44419	Other building material dealers
Aggregate	4461	Health and personal care stores
Disaggregate	44611	Pharmacies and drug stores
	44612	Optical goods stores
	44613	Other health and personal care stores
	44619	Cosmetics, beauty supplies, and perfume stores

4.4 Simple Examples

This section concludes with a more detailed presentation of the uncertainty and correlation data of particular accounting identities, which illustrate the different qualitative patterns described in Section 3.4.

Four accounting identities were chosen, each with four disaggregate data points, whose properties are summarized in Tables 1–6. Table 1 summarizes the properties of the aggregate industries, Tables 2–5 summarize the properties of disaggregate industries, and Table 6 identifies disaggregate data.

Table 1 shows that in the first accounting identity, whose disaggregate data is described in Table 2, aggregate uncertainty exceeds the zero-correlation uncertainty, while in the remainder accounting identities, whose disaggregate data is described in Tables 3–5, aggregate uncertainty is below the zero-correlation uncertainty.

Furthermore, Table 1 also shows that, as expected, the maximum-entropy aggregate uncertainty is always larger than the zero-correlation uncertainty, but always closer to the latter than to the maximum value. Tables 2–5 illustrate the pattern that the highest correlations (in absolute terms) are found between industries with the largest employment uncertainty.

In Table 3 the lower bound is positive, in which case it is expected that if aggregate uncertainty drops to the lower bound, then all correlations become plus or minus one.

In Table 4 there is still a large distance between the largest disaggregate industry employment uncertainty and the next figure, but the lower bound is now zero. In this case, although there are both positive and negative correlations, they will not rise or fall to plus or minus one if the aggregate correlation becomes zero.

Finally, in Table 5 the largest disaggregate industry employment uncertainty is very close to the other disaggregate uncertainties. In this case, no matter how low aggregate uncertainty drops, all correlations remain negative.

Table 6 presents the NAICS codes and description of the disaggregate data in the selected accounting identities.

5. CONCLUSIONS

This article applies concepts and tools from Bayesian inference, and in particular the MEP, to determine the prior probability distribution, uncertainty, and correlations of statistical economic data, when additional information such as best guesses and accounting identities are available.

The main findings of this article are:

1. The prior probability distribution of a statistical datum of which a best guess and uncertainty estimate are known is a truncated Gaussian.
2. The prior relative uncertainty of an isolated datum of which only a best guess is known is unitary.
3. The prior correlation of data connected through an accounting identity can be determined by solving Equation (3.8), and there are both a lower and an upper bound to aggregate uncertainty.
4. If the aggregate best guess is not known, then disaggregate data are uncorrelated whereas if the aggregate best guess is known but aggregate uncertainty is not, then all prior correlations are positive.
5. If the aggregate uncertainty is known, prior correlations can be either all positive, all negative, or a mix of both, depending on the relative values of aggregate and disaggregate uncertainties.

These results represent an important contribution to the existing literature on the uncertainty of statistical economic data, by identifying under which conditions a particular probability distribution and set of uncertainties and correlations should be used. In particular, the conventional assumptions of nontruncated Gaussian and zero correlations were shown to be particular cases of a more general framework corresponding, respectively, to the situation of low relative uncertainty and absence of an independent estimate for aggregate data.

In this study, the theoretical results were complemented by the estimation of the uncertainties and correlations of an empirical dataset, the CBP database, in which a wide range of values of both relative uncertainties and correlations was found, with many uncertainties outside the range in which the nontruncated Gaussian is acceptable and many correlations being substantially different from zero. Caution in generalizing these results is required, as the distribution of data varies significantly across counties (depending on the size and the nature of its economy) and over time.

An important direction of future research that would complement the present study is the generalization of the expression for the empirical determination of prior correlations, Equation (3.8), from the nontruncated to the truncated Gaussian multivariate case.

Another interesting open question is the derivation of a concentration theorem (Jaynes 1979), and the clarification of how

large is the entropy of the MEP solution relative to other consistent solutions, and eventually to other priors (Dias and Shimony 1981; Gokhale and Press 1982; Uffink 1995; Kass and Wasserman 1996; Fernandez-Alcala, Navarro-Moreno, and Ruiz-Molina 2007; Rodriguez and Horst 2008; Sanso, Forest, and Zantedeschi 2008; Huang and Wand 2013).

The computational challenges that lie ahead should not be underestimated. The application of the two-stage Newton method developed here to large and complex datasets is nontrivial and requires further refinement and optimization.

Finally, the priors derived here are worst-case solutions to which a practitioner should fall back in the absence of better information. If the practitioner has expert knowledge suggesting that other priors are a better description of the system being studied, these alternative priors should be used for as long as they are properly justified and mutually consistent.

SUPPLEMENTARY MATERIALS

Supplementary materials contain Appendices A–F.

ACKNOWLEDGMENTS

This article benefitted from comments by many colleagues and friends. I especially thank Hai Xiang Lin for his thorough review and both Michael Lahr and Sumei Zhang for sharing data for comparison. Any errors it may contain are the author's sole responsibility.

[Received April 2014. Revised March 2015.]

REFERENCES

Abowd, J. M., and Vilhuber, L. (2005), "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers," *Journal of Business & Economic Statistics*, 23, 133–152. [363]

Antonello, P. (1990), "Simultaneous Balancing of Input-Output Tables at Current and Constant Prices With First Order Vector Autocorrelated Errors," *Economic Systems Research*, 2, 157–172. [361]

Ballantyne, A. P., Alden, C. B., Miller, J. B., Tans, P. P., and White, J. W. C. (2012), "Increase in Observed Net Carbon Dioxide Uptake by Land and Oceans During the Past 50 Years," *Nature*, 488, 70–72. [361]

Castillo, J. D. (1994), "The Singly Truncated Normal Distribution: A Non-Steep Exponential Family," *Annals of the Institute of Statistical Mathematics*, 46, 57–66. [359]

Chen, B. (2012), "A Balanced System of U.S. Industry Accounts and Distribution of the Aggregate Statistical Discrepancy by Industry," *Journal of Business & Economic Statistics*, 30, 202–211. [357,362]

Cholette, P., and Dagum, E. (1994), "Benchmarking Time Series With Autocorrelated Sampling Errors," *International Statistics Review*, 62, 365–377. [361]

Chow, G. C., and Lin, A. L. (1971a), "Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series," *Journal of the American Statistical Association*, 71, 719–721. [361]

——— (1971b), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," *Review of Economics and Statistics*, 53, 372–375. [361]

Clemen, R. T., and Winkler, R. L. (1985), "Limits for the Precision and Value of Information From Dependent Sources," *Operations Research*, 33, 427–442. [357]

Cohen, A. C. (1950), "Estimating the Mean and Variance of Normal Populations From Singly Truncated and Doubly Truncated Samples," *Annals of Mathematical Statistics*, 21, 557–569. [359]

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory*, New York: Wiley. [359]

Cunningham, A., Eklund, J., Jeffery, C., Kapetanios, G., and Labhard, V. (2012), "A State Space Approach to Extracting the Signal From Uncertain Data," *Journal of Business & Economic Statistics*, 30, 173–180. [357]

Dagum, E. B., and Cholette, P. A. (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, New York: Springer-Verlag. [357,361]

Dagum, E. B., Cholette, P. A., and Chen, Z.-G. (1998), "A Unified View of Signal Extraction, Benchmarking, Interpolation and Extrapolation of Time Series," *International Statistical Review*, 66, 245–269. [361]

Dias, P. M. C., and Shimony, A. (1981), "A Critique of Jaynes' Maximum Entropy Principle," *Advances in Applied Mathematics*, 2, 172–211. [366]

Díaz, B., and Morillas, A. (2011), "Incorporating Uncertainty in the Coefficients and Multipliers of an IO Table: A Case Study," *Papers in Regional Science*, 90, 845–861. [358]

Dietzenbacher, E. (2006), "Multiplier Estimates: To Bias or Not to Bias?" *Journal of Regional Science*, 46, 773–786. [357,358,361]

Dowson, D., and Wragg, A. (1973), "Maximum-Entropy Distributions Having Prescribed First and Second Moments (Correspondence)," *IEEE Transactions on Information Theory*, 19, 689–693. [359]

Doyle, P., Lane, J. I., Theeuwes, J. J. M., and Zayatz, L. V. (2001), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam, The Netherlands: Elsevier B. V. [363]

Engle, R., and Kelly, B. (2012), "Dynamic Equicorrelation," *Journal of Business & Economic Statistics*, 30, 212–228. [361]

Fernandez-Alcala, R. M., Navarro-Moreno, J., and Ruiz-Molina, J. C. (2007), "Functional Estimation Incorporating Prior Correlation Information," *Computational Statistics*, 22, 439–447. [366]

Fischetti, M., and Salazar, J. J. (2005), "A Unified Mathematical Programming Framework for Different Statistical Disclosure Limitation Methods," *Operations Research*, 53, 819–829. [362]

Flugsrud, K., and Hoem, B. (2011), "Uncertainties in the Norwegian Greenhouse Gas Emission Inventory," Report 35/2011, Statistics Norway, Oslo, Norway. [361]

Frank, S. A., and Smith, D. E. (2010), "Measurement Invariance, Entropy, and Probability," *Entropy*, 12, 289–303. [359]

Gokhale, D. V., and Press, S. J. (1982), "Assessment of a Prior Distribution for the Correlation Coefficient in a Bivariate Normal Distribution," *Journal of the Royal Statistical Society*, Series A, 145, 237–249. [366]

Golan, A., Judge, G., and Robinson, S. (1994), "Recovering Information From Incomplete or Partial Multisectoral Economic Data," *Review of Economics and Statistics*, 76, 541–549. [358,361]

Golan, A., and Vogel, S. J. (2000), "Estimation of Non-Stationary Social Accounting Matrix Coefficients With Supply-Side Information," *Economic Systems Research*, 12, 447–471. [358]

Greene, W. H. (2008), *Econometric Analysis* (6th ed.), Upper Saddle River, NJ: Prentice Hall. [359]

Guldmann, J.-M. (2013), "Analytical Strategies for Estimating Suppressed and Missing Data in Large Regional and Local Employment, Population, and Transportation Databases," *Wiley Interdisciplinary Reviews—Data Mining and Knowledge Discovery*, 3, 280–289. [357,362]

Heijungs, R., and Suh, S. (2002), *The Computational Structure of Life Cycle Assessment*, Dordrecht, The Netherlands: Kluwer Academic. [360]

Hendry, D. F., and Hubrich, K. (2011), "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate," *Journal of Business & Economic Statistics*, 29, 216–227. [358]

Huang, A., and Wand, M. P. (2013), "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8, 439–452. [366]

Hyslop, D. R., and Imbens, G. W. (2001), "Bias From Classical and Other Forms of Measurement Error," *Journal of Business & Economic Statistics*, 19, 475–481. [357]

Isserman, A. M., and Westervelt, J. (2006), "1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data," *International Regional Science Review*, 29, 311–335. [362]

Jaynes, E. T. (1957), "Information Theory and Statistical Mechanics I," *Physical Review*, 106, 620–630. [358]

——— (1979), "Concentration of Distributions at Entropy Maxima," in E. T. Jaynes: *Papers on Probability, Statistics and Statistical Physics* (book published in 1989), ed. R. D. Rosenkrantz, Dordrecht, The Netherlands: Kluwer Academic, pp. 315–336. [365]

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press. [357,358]

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions* (Vol. 1), New York: Wiley. [359]

Kass, R. E., and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370. [366]

Lahr, M. L., and Mesnard, L. d. (2004), "Biproportional Techniques in Input-Output Analysis: Table Updating and Structural Analysis," *Economic Systems Research*, 16, 115–134. [363]

Lahr, M. L., and Stevens, B. H. (2002), "A Study of the Role of Regionalization in the Generation of Aggregation Error in Regional Input-Output Models," *Journal of Regional Science*, 42, 477–507. [362]

Lenzen, M. (2001), "Errors in Conventional and Input-Output-Based Life-Cycle Inventories," *Journal of Industrial Ecology*, 4, 127–148. [361]

——— (2011), "Aggregation Versus Disaggregation in Input-Output Analysis of the Environment," *Economic Systems Research*, 23, 73–89. [361]

Lenzen, M., Wood, R., and Wiedmann, T. (2010), "Uncertainty Analysis for Multi-Region Input-Output Models: A Case Study of the UK's Carbon Footprint," *Economic Systems Research*, 22, 43–63. [358]

Manski, C. F. (2014), "Communicating Uncertainty in Official Economic Statistics," Working Paper 20098, National Bureau of Economic Statistics, Cambridge, MA. [357]

Meijer, E., Rohwedder, S., and Wansbeek, T. (2012), "Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data," *Journal of Business & Economic Statistics*, 30, 191–201. [357]

Miller, R. E., and Blair, P. D. (2009), *Input-Output Analysis: Foundations and Extensions* (2nd ed.), Cambridge: Cambridge University Press. [357]

Nicoletti, C., Peracchi, F., and Foliano, F. (2011), "Estimating Income Poverty in the Presence of Missing Data and Measurement Error," *Journal of Business & Economic Statistics*, 29, 61–72. [357]

Pearson, K., and Lee, A. (1908), "On the Generalized Probable Error in Multiple Normal Correlations," *Biometrika*, 24, 55–64. [359]

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007), *Numerical Recipes: The Art of Scientific Computing* (3rd ed.), New York: Cambridge University Press. [361]

Rampa, G. (2008), "Using Weighted Least Squares to Deflate Input-Output Tables," *Economic Systems Research*, 20, 259–276. [361]

Rassier, D., Howells, T., Morgan, E., Empey, N., and Roesch, C. (2007), "Implementing a Reconciliation and Balancing Model in the U.S. Industry Accounts," Working Paper WP2007-4, U.S. Bureau of Economic Analysis. [361]

Rodrigues, J. F. D. (2014), "A Bayesian Approach to the Balancing of Statistical Economic Data," *Entropy*, 16, 1243–1271. [359,363]

Rodriguez, A., and Horst, E. t. (2008), "Bayesian Dynamic Density Estimation," *Bayesian Analysis*, 3, 339–366. [366]

Rypdal, K., and Zhang, L.-C. (2000), "Uncertainties in the Norwegian Greenhouse Gas Emission Inventory," Report 13/2000, Statistics Norway, Oslo, Norway. [361]

Sanso, B., Forest, C. E., and Zantedeschi, D. (2008), "Inferring Climate System Properties Using a Computer Model," *Bayesian Analysis*, 3, 1–38. [366]

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27, 379–423. [358]

Stone, R., Champernowne, D. G., and Meade, J. E. (1942), "The Precision of National Income Estimates," *Review of Economic Studies*, 9, 111–125. [357]

Tallis, G. M. (1961), "The Moment Generating Function of the Truncated Multi-Normal Distribution," *Journal of the Royal Statistical Society*, Series B, 23, 223–229. [359]

Ten Raa, T., and Rueda-Cantuche, J. M. (2003), "The Construction of Input-Output Coefficients Matrices in an Axiomatic Context: Some Further Considerations," *Economic Systems Research*, 15, 439–455. [357]

Treyz, G. I., and Stevens, B. H. (1985), "The TFS Regional Modelling Methodology," *Regional Studies*, 19, 547–562. [362]

Uffink, J. (1995), "Can the Maximum Entropy Principle be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Modern Physics*, 26, 223–261. [366]

Weale, M. (1988), "The Reconciliation of Values, Volumes and Prices in National Accounts," *Journal of the Royal Statistical Society*, Series A, 151, 211–221. [361]

Weise, K., and Woger, W. (1992), "A Bayesian Theory of Measurement Uncertainty," *Measurement Science and Technology*, 4, 1–11. [357,359]

Wood, R. (2009), "Construction, Stability and Predictability of an Input-Output Time-Series for Australia," *Economic Systems Research*, 23, 175–211. [357]

Zhang, S., and Guldmann, J.-M. (2009), "Estimating Suppressed Data in Regional Economic Databases: A Goal-Programming Approach," *European Journal of Operations Research*, 192, 521–537. [362]

Zhang, S., and Guldmann, J.-M. (2015), "A Regression-Constrained Optimization Approach to Estimating Suppressed Information Using Time-Series Data: Application to County Business Patterns 1999–2006," *International Regional Science Review*, 38, 119–150. [362]