Maximum-Entropy Prior Uncertainty and Correlation of Statistical Economic Data: Supplementary Information

March 9, 2015

A Conversion of truncated Gaussian parameters

Let T be a truncated Gaussian distribution, $T > 0$, with Gaussian parameters $\tilde{m}$ and $\tilde{s}$ and observables m and s (best guess and uncertainty), and let $u = s/m$ and $\tilde{u} = \tilde{m}/\tilde{s}$. Let $\phi(y)$ and $\Phi(y)$ denote the probability density function and the cumulative distribution function of the standard normal distribution.

According to Greene (2008), the connection between the Gaussian parameters and the observables is:

$$m = \tilde{m} + \tilde{s}\frac{\phi\left(-\tilde{u}\right)}{1-\Phi\left(-\tilde{u}\right)}; \tag{A.1}$$

$$s^2 = \tilde{s}^2\left(1 - \frac{\phi\left(-\tilde{u}\right)}{1-\Phi\left(-\tilde{u}\right)}\left(\tilde{u} + \frac{\phi\left(-\tilde{u}\right)}{1-\Phi\left(-\tilde{u}\right)}\right)\right). \tag{A.2}$$

In the previous expression the ratio $\phi(-\tilde{u})/(1-\Phi(-\tilde{u}))$ is known as the inverse Mills ratio. Expressions with different degrees of accuracy for $\Phi(-\tilde{u})$ can be found in Abramowitz and Stegun (1964). Equations. A.1 and A.2 are no longer accurate when $\tilde{u} < -5$ (corresponding to $u > 0.98$), and in this region, a better approximation is given by:

$$m = \frac{\tilde{s}^2}{\tilde{m}}\left(-1 + 1.97777\left(1-u\right) + 6.8\left(1-u\right)^2\right); \tag{A.3}$$

$$s = m\left(1 - 0.014172\left(\tilde{u} + \sqrt{16.8 + \tilde{u}^2}\right)^2\right). \tag{A.4}$$

The author is unaware of any analytical formula which expresses the Gaussian parameters as a function of observables. By inspection, the following set of conversion functions was obtained. If $u \leq 0.3$, then $\tilde{m} = m$ and $\tilde{s} = s$. If $u > 0.3$, then $\tilde{m} = m g_m^{tp}(u)$, while:

$$\tilde{s} = \begin{cases} \dfrac{s}{\sqrt{g_s^{tp}(u)}} & \text{if } 0.3 < u \leq 0.8; \\ -s\dfrac{g_m^{tp}(u)}{\sqrt{g_m^{tp}(u) g_r^{tp}(u)}} & \text{if } u > 0.8. \end{cases}$$

The conversion functions are:

i	c_i^m	c_i^g	c_i^s	c_i^r
1	0.937492	0.468838	0.546626	0.83179
2	1.78863	0.118555	0.439319	0.617251
3	7.13173	0.00235939	1.83447	6.3836
4	5.42261			

Table 1: Coefficients of conversion between observable and Gaussian parameters.

$$g_m^{tp}(u) = 1 - \frac{1}{1-u} e^{-(1-u)\left(c_4^m + c_3^m |u - c_1^m|^{c_2^m}\right)} + \frac{c_3^g}{c_2^g} \phi \left(\frac{u - c_1^g}{c_2^g} \right); \tag{A.5}$$

$$g_s^{tp}(u) = (1-u) \frac{c_3^s}{c_2^s} \phi \left(\frac{u - c_1^s}{c_2^s} \right); \tag{A.6}$$

$$g_r^{tp}(u) = (1-u) \frac{c_3^r}{c_2^r} \phi \left(\frac{u - c_1^r}{c_2^r} \right) - 1. \tag{A.7}$$

Parameters c_i^* are reported in Table 1. Equations. A.1-A.7 allow the conversion from observables to parameters and vice-versa, keeping the relative error between initial and final observables below 0.5% in the whole range $0 < u < 1$.

B Multivariate MEP solution

The derivation of the analytical solution of correlation priors constrained by Eq. 1.1 when all uncertainties of both disaggregate and aggregate data are known is as follows. The Lagrangian reads:

$$
\begin{aligned}
L = & \int_\Omega dq\, p(\mathbf{q}) \ln\left(p(\mathbf{q})\right) + \int_0^\infty dq_0\, p(q_0) \ln\left(p(q_0)\right) \\
& + \lambda \left(\int_\Omega dq\, p(\mathbf{q}) - 1 \right) + \lambda_0 \left(\int_0^\infty dq_0\, p(q_0) - 1 \right) \\
& + \sum_{i=0}^{n} \alpha_i \left(\mathrm{E}[t_i] - m_i \right) + \sum_{i=0}^{n} \beta_i \left(\mathrm{E}[t_i^2] - \mathrm{E}[t_i]^2 - s_i^2 \right) \\
& + \beta \left(\sum_{i=1}^{n} \left(\mathrm{E}[t_i^2] - \mathrm{E}[t_i]^2 - s_i^2 \right) - \left(\mathrm{E}[t_0^2] - \mathrm{E}[t_0]^2 - s_0^2 \right) \right. \\
& \left. + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(\int_\Omega dq\, p(\mathbf{q}) q_i q_j - m_i m_j \right) \right) .
\end{aligned} \tag{B.1}
$$

In Eq. B.1 the expression $\int_\Omega dq$ is a shorthand for the product $\prod_{i=1}^{n} \int_0^\infty dq_i$. Each q_i is the realization of the random variable t_i. The first term on the right hand side of Eq. B.1 is the entropy of the joint probability distribution of the disaggregate data and the second term is the probability of the aggregate datum. The second line contains the normalization constraints. The third line contains the best guess and uncertainty constraints (recall that, for the time being, we assume they are known). The fifth and sixth lines contain the constraint on second-order moments, Eq. 3.2.

Minimization of Eq. B.1 with respect to $p(\mathbf{q})$ yields:

$$
0 = -(\ln p(\mathbf{q}) + 1) + \lambda + \sum_{i=1}^{n} \alpha_i q_i + \sum_{i=1}^{n} \beta_i q_i^2 + 2\beta \sum_{i=2}^{n} \sum_{j=1}^{i-1} q_i q_j + C.
$$

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

$$
p(\mathbf{q}) = C \exp\left(\sum_{i=1}^{n} \beta_i q_i^2 + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \beta q_i q_j + \sum_{i=1}^{n} \alpha_i q_i \right) . \tag{B.2}
$$

Notice that the exponent in Eq. B.2 is a second-degree polynomial whose coefficients are Lagrange multipliers. The exponent of the probability density function of the multivariate Gaussian is also a second-degree polynomial:

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

which can be expanded as:

$$\begin{aligned} p(\mathbf{q}) = C_1 \exp\Bigg(& -\sum_{i=1}^{n} \frac{(\tilde{s}^{-1})_{ii}}{2} q_i^2 - 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \frac{(\tilde{s}^{-1})_{ij}}{2} q_i q_j \\ & +2\sum_{i=1}^{n}\left(\sum_{j=1}^{n} \frac{(\tilde{s}^{-1})_{ij}}{2}\tilde{m}_j\right) q_i + C_2\Bigg). \end{aligned} \tag{B.3}$$

In the previous expression $(\tilde{s}^{-1})_{ij}$ is the (i,j) entry of the inverse covariance matrix $\tilde{\mathbf{S}}^{-1}$. The coefficients which precede either q_i^2 or q_i are not important, since all best guesses and uncertainties are assumed to be known. Comparison between the term preceding product $q_i q_j$ in Eq. B.2 and in Eq. B.3 leads to the solution to determine prior correlations: *all prior inverse covariances are identical.* If the covariance matrix is expanded as $\tilde{\mathbf{S}} = \mathrm{diag}(\tilde{\mathbf{s}})\tilde{\mathbf{R}}(\mathrm{diag}\,\tilde{\mathbf{s}})$, where diag is a diagonal matrix, $\tilde{\mathbf{s}}$ is the uncertainty vector and $\tilde{\mathbf{R}}$ is the correlation matrix, then the constraint can be simplified to Eq. 3.6.

If the uncertainty of the aggregate datum is unknown, then in the second term of the third line of Eq. B.1 the iterator should start at 1 rather than 0, removing the constraint on aggregate uncertainty. Minimization of Eq. B.1 with respect to $p(q_0)$ now yields:

$$0 = -(\ln p(q_0) + 1) + \lambda_0 + \alpha_0 q_0 - \beta q_0^2.$$

Recalling the form of the univariate Gaussian distribution, Eq. 2.4, the prior aggregate uncertainty is given by Eq. 3.7.

C Correlation and entropy

The effect of correlations on the joint probabililty of the disaggregate data, $\sum_{i=1}^{n} t_i$, is as follows. The entropy of a multivariate Gaussian is (Cover and Thomas, 1991):

$$L_1 = \frac{1}{2}\left(n \log(2\pi e) + 2\sum_{i=1}^{n} \log(s_i) + \log(\det(\mathbf{R})) \right). \qquad \text{(C.1)}$$

Thus, the configuration of correlations which maximizes entropy is also that which maximizes the determinant of the correlation matrix. For simplicity, consider that all correlations are identical, $r_{ij} = r$. If this is the case, it is possible to apply Sylvester's theorem (Akritas et al., 1996; Saled and Said, 2008):

$$\det(\mathbf{X} + \mathbf{u}\mathbf{v}') = \det(\mathbf{X}) \det(1 + \mathbf{v}'\mathbf{X}^{-1}\mathbf{u}),$$

where $\mathbf{X}$ is a matrix and both $\mathbf{u}$ and $\mathbf{v}$ are column vectors. If $\mathbf{e}$ denotes a column vector of ones and $\mathbf{R}$ is expressed as $(1-r)\mathbf{I} + (\sqrt{r}\mathbf{e})(\sqrt{r}\mathbf{e})'$, the theorem implies that:

$$\det(\mathbf{R}) = \det((1-r)\mathbf{I}) \det\left(1 + n\frac{r}{1-r}\right) = (1-r)^{n-1}(1+(n-1)r).$$

Substituting this information in Eq. C.1 leads to:

$$L_1 = \frac{1}{2}\left((n-1)\log(1-r) + \log(1+(n-1)r) + \ldots\right),$$

where $\ldots$ are terms independent of r. The correlation which minimizes entropy is found by setting the derivative of entropy with respect to r equal to zero:

$$\frac{dL_1}{dr} = \frac{n-1}{2}\left(\frac{-1}{1-r} + \frac{1}{1+(n-1)r}\right) = 0,$$

from where it follows that $r = 0$. This result is not surprising: the more uncorrelated a set of variables is, the larger is the set of possible combinations.

The link between the correlations among disaggregate data and the uncertainty of the aggregate datum is as follows. If the latter is described by a nontruncated Gaussian, $s_0/m_0 < 0.3$, its entropy is:

$$L_0 = \log(s_0) + \frac{1}{2}\log(2\pi e).$$

If the aggregate datum is described by an exponential, $s_0/m_0 = 1$, its entropy is:

$$L_0 = \log(s_0) + 1.$$

In the intermediate case, $0.3 < s_0/m_0 < 1$, no analytical solution exists but the following numerical approximation yields an error which is no larger than 2%:

$$L_0 = \log(s_0) + \left(1 - \frac{s_0}{m_0}\right)^3 \frac{1}{2}\log(2\pi e) + \left(\frac{s_0}{m_0}\right)^3 .$$

According to Eq. 3.2 the aggregate uncertainty is a monotonically increasing function of correlations, $\partial s_0/\partial r_{ij} > 0$. Because the entropy of the aggregate datum, L_0, is itself a monotonically increasing function of aggregate uncertainty, $\partial L_0/\partial s_0 > 0$, it follows that the set of correlations which maximizes the entropy of the aggregate datum is one, $r = 1$.

D Qualitative patterns

The following general patterns can be observed, concerning the behaviour of disaggregate correlations as a function of disaggregate uncertainties:

1. In absolute terms, the correlation between a pair of larger uncertainties is larger than the correlation between a pair of smaller uncertainties, i.e., if $s_1 > s_2 > s_3$ then $|r_{12}| > |r_{13}|$.
2. If aggregate uncertainty is maximal, $s_0 = s_{\text{max}}$, then all disaggregate data are perfectly correlated, i.e., $r_{ij} = 1$ with $i > 0$ and $j > 0$.
3. If aggregate uncertainty is larger than the zero-correlation uncertainty, $s_0 > s_{\text{zero}}$, then all correlations are positive, i.e., $r_{ij} > 0$ with $i > 0$ and $j > 0$.

4. If aggregate uncertainty is minimal, $s_0 = s_{\min}$, and minimal uncertainty is positive, $s_{\min} > 0$, then the largest disaggregate datum is perfectly anti-correlated with all other data, and the other data are perfectly correlated, i.e., if $s_1 > s_i$ with $i > 1$, then $r_{1i} = -1$ and $r_{ij} = 1$ with $j > 1$.

5. If aggregate uncertainty is much smaller than the zero-correlation uncertainty and the largest disaggregate datum is much larger than the remainder disaggregate data where $i > 1$, then the correlation between the largest disaggregate datum and the remainder are negative, $r_{1i} < 0$, whereas the correlations among the remainder are positive, $r_{ij} > 0$ with $j > 1$ and s_1 being the largest disaggregate uncertainty. Naturally, what defines “much smaller” and “much larger” depends on the particular configuration of uncertainties.

6. In all other cases in which aggregate uncertainty is smaller than the zero correlation uncertainty, $s_0 < s_{\text{zero}}$, all correlations are negative, $r_{ij} < 0$ with $i > 0$ and $j > 0$.

E Data balancing

The best guess estimates of the empirical application were balanced using the following algorithm.

Initial best guesses and relative uncertainties were arranged in vectors $\mathbf{m}(0)$ and $\mathbf{u}$, respectively, and accounting identities were arranged in matrix $\mathbf{G}$ where $G_{ij} = -1$ if in accounting identity i entry j is the aggregate datum, $G_{ij} = 1$ if entry j is a disaggregate datum and $G_{ij} = 0$ otherwise. Additionally, $\mathbf{f}$ is a disclosure flag vector which is $f_i = 1$ if the data point is disclosed and $f_i = 0$ otherwise.

Nondisclosed data is adjusted while disclosed data remains fixed. A set of constraints and truncated aggregation matrix are obtained as:

$$\begin{gathered}\mathbf{k} = \mathbf{G}\left(\mathbf{f}\#\mathbf{m}\right);\\ \mathbf{G}^* = \mathbf{G}\,\mathrm{diag}(\mathbf{i} - \mathbf{f}),\end{gathered}$$

where # is entry-wise (Hadamard) product and $\mathbf{i}$ is a vector of ones. Iteration then proceeds as:

$$\begin{aligned}\boldsymbol{\alpha}(i) &= -\left(\mathbf{G}^*\mathbf{m}(i) + \mathbf{k}\right) \div \mathrm{diag}\left(\mathbf{G}^*\,\mathrm{diag}(\mathbf{u}\#\mathbf{m}(i))(\mathbf{G}^*)'\right);\\ \mathbf{m}(i+1) &= \mathbf{m}(i) + c(i)\,\mathrm{diag}(\mathbf{u}\#\mathbf{m}(i))(\mathbf{G}^*)'\boldsymbol{\alpha}(i),\end{aligned}$$

where $\div$ is entry-wise (Hadamard) division and $c(i)$ is largest real number in the range $(0, 1)$ such that:

$$\max\left\{|m_j(i+1) - m_j(i)|/m_j(i)\right\}_j < c_{\max},$$

where $c_{\max} = 0.1$ is the maximum allowed relative displacement.

The data balancing process finishes when the largest error per constraint falls below the threshold of $\epsilon = 0.1$ jobs, $\max\left\{|d_j(i)|\right\}_j < \epsilon$, where the error of every constraint is obtained as:

$$\mathbf{d}(i) = \mathbf{G}^*\mathbf{m}(i) + \mathbf{k}.$$

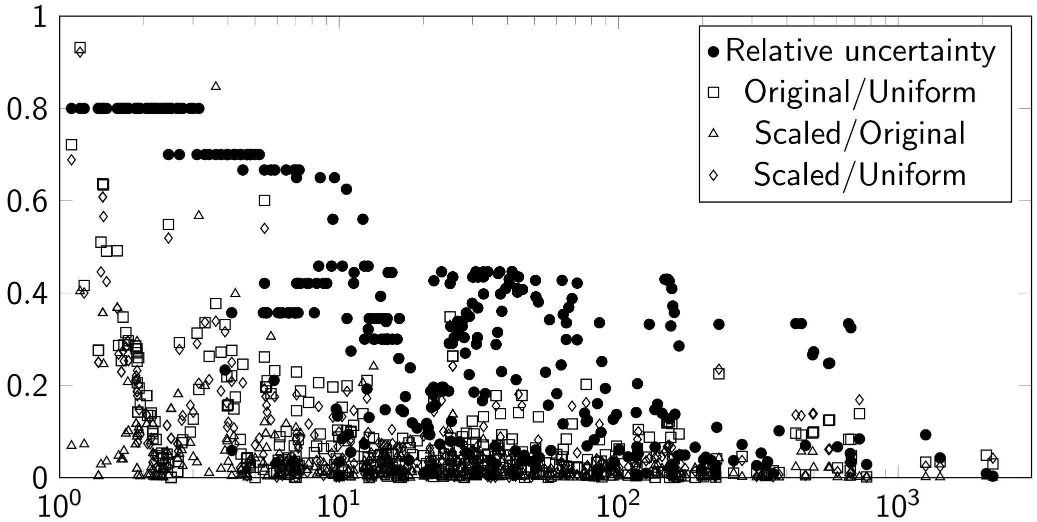

Figure 1: Absolute relative distance between best guess estimates, using different choices of prior (original or scaled) and balancing method (original or uniform).

F Comparison of processing procedures

In order to assess the robustness of the empirical findings several variations were on the processing procedure described in Section 4.2 were performed.

Besides the original prior configuration of best guesses, an alternative configuration was considered in which every best guess prior was scaled down by a factor of 0.958. This alternative prior was chosen because it was observed that using the original prior the prior discrepancies in accounting identities were positive on average. That is, the average nondisclosed best guess was being over-estimated. The above mentioned factor was chosen such as to minimize the average discrepancy in accounting identities.

Besides the original iterative weighted least squares balancing algorithm, in which relative uncertainties were used as weights, an alternative balanc-

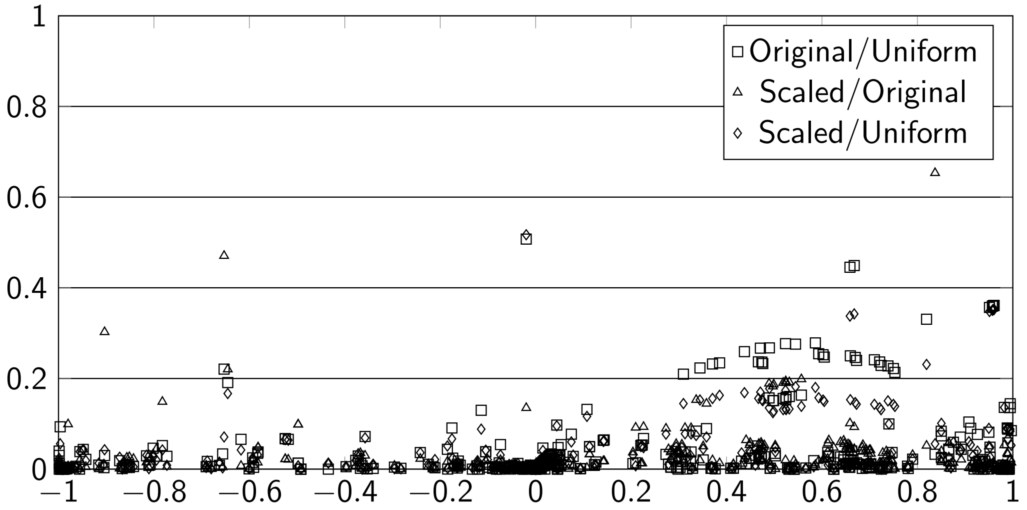

Figure 2: Absolute distance between correlation estimates, using different choices of prior (original or scaled) and balancing method (original or uniform).

ing method was considered. This alternative balancing method was still an iterative least squares method but now it was ordinary rather than weighted. That is, all weighting adustments were considered to be identical, irrespective of the relative uncertainty of the datum.

Figure 1 shows the absolute relative distances between the best guess estimates used in Section 4 and those that are obtained using the different combinations of the original or scaled prior and original or uniform weighed balancing. This figure shows that most absolute relative distances fall below the relative uncertainty of the data.

Figure 2 shows the absolute distances between the correlation estimates obtained in Section 4 and those that would be obtained using the different combinations of the original or scaled prior and original or uniform weighed balancing. This figure shows that most absolute distances are below 0.1. The only region of the state space where distances are systematically larger is in the positive range from 0.2 to 0.8.

In general, the original data is more similar to data using the scaled prior and the original balancing than to data using uniform weights in balancing.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* National Bureau of Standards, Washignton DC, USA, 1964.

A. G. Akritas, E. K. Akritas, and G. I. Malaschonok. Various proofs of Sylvester's (determinant) identity. *Mathematics and Computers in Simulation*, 42 (4-6):585–596, 1996.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley and Sons, Inc., New York, 1991.

W. H. Greene. *Econometric Analysis.* Prentice Hall, Upper Saddle River, NJ, 2008. 6th Edition.

A. Saled and K. Said. A simple proof of Sylvester's (determinants) identity. *Applied Mathematical Sciences*, 32 (2):1571–1580, 2008.