Short Communication

Global Correlation and Uncertainty Accounting

Roger M. Cooke1*, Sassan Saatchi2 and Stephen Hagen3

1Resources for the Future, USA
2Jet Propulsion Lab, California University of Technology, USA
3Applied Geosolutions, USA

Abstract

For a high dimensional field of random variables, global correlation is defined as the ratio of average covariance and average variance, and its elementary properties are studied. Global correlation is used to harmonize uncertainty assessments at global and local scales. It can be estimated by the correlation of random aggregations of fixed size of disjoint sets of random variables. Illustrative applications are given using crop loss per county per year and forest carbon.

INTRODUCTION

This note defines global correlation, studies its elementary properties and illustrates its use in global uncertainty accounting. The rich literature on multivariate correlation can receive only passing mention in this space constrained note. Conical correlation [1] concerns the maximal product moment correlation between linear combinations of two random vectors; interclass correlation [2] describes the correlations in grouped data. Multiple correlation and the correlation ratio [3] relate a single variable to a set of variables. Random correlation matrices [4] and the distribution of their determinate [5] have sparked interest in the (scaled) determinant of the correlation matrix as a measure of multivariate association. Using Vines [6-9] have made progress in understanding random determinants of correlation matrices. Unfortunately, these techniques are ill suited to the problems encountered here; harmonizing uncertainty assessment in aggregations of up to 4 billion variables.

METHODS

The correlation of random aggregates is used to estimate global correlation. The following facts and definitions are used:

1) If $X_1, X_2$ are independent and identically distributed random variables with standard deviation $\sigma$, then

$$E(X_1 - X_2)^2 = 2\sigma^2$$

(1)

2) If $X_1, \ldots, X_N$ have average variance $\sigma^2$ and average covariance $c$ then

$$VAR(\Sigma X_i) = \sigma^2 N + N(N - 1)c$$

(this shows that $c \geq -\sigma^2 / (N - 1)$, since LHS $\geq 0$). (2)

3) Define $\rho = c / \sigma^2$ as the global correlation. If $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$ have average covariance $c$ and average variance $\sigma^2$, both within and between components, then

$$\rho(\Sigma_{i=1}^N X_i, \Sigma_{i=1}^N Y_i) = N^2 c / (N\sigma^2 + N (N - 1)c) = N \rho / (1 + (N - 1)\rho).$$

(3)

The correlation of sums of $N$ converges to 1 as $N \to \infty$, for any $\rho > 0$. If $\rho > 0$ and $N >> 1$, then

$$StDev(\Sigma X_i) = \sigma N(N^{-1} + \rho(N - 1))^{-1/2} \approx \sigma N^{1/2}. \tag{4}$$

This should be compared to the case where $c = 0$, which holds if the $X_i$ are independent:

$$StDev(\Sigma X_i) = \sigma N^{1/2}. \tag{5}$$

With independence, the uncertainty (standard deviation) of the sum grows with $N^{1/2}$, but a small global correlation $\rho$ causes the growth to be linear in $N$. To appreciate this, let $\rho$ be the global correlation of the amount of forest carbon per hectare; we wish to assess the uncertainty of global forest carbon based on the average variance in the estimates per hectare. The number of hectares of forest on the earth is $N = 4E9$, $N^{1/2} = 63,246$, etc.
whereas with \( \rho = 0.001, \) \( \rho_0 = 4,000,000. \) The difference between the cases \( \rho = 0 \) and \( \rho = 0.001 \) is huge. Recall:

**Cauchy Schwarz Inequality:**

\[
X, Y \in \mathbb{R}^N, \ (\Sigma x_i y_i)^2 \leq \Sigma x_i^2 \Sigma y_i^2.
\]

Equality holds if and only if \( y_i = A x_i \) for some \( A \in \mathbb{R} \). Take \( y_i = 1 \), then \( (\Sigma x_i)^2 \leq N \Sigma x_i^2 \) with equality if and only if the \( x_i \) are constant. Equivalently, if \( x = \text{average} (x_i) \): \( (N x)^2 \leq N \Sigma x_i^2 \) or \( x^2 \leq \Sigma x_i^2 / N \) (a version of Jensen's inequality) with equality if and only if the \( x_i \) are constant

**RESULTS AND DISCUSSION**

**Lemma 1:** For all \( X \in \mathbb{R}^N, (N-1) \Sigma x_i^2 \geq \Sigma_{i \neq k} x_i x_k \).

**Proof:**

Put

\[
\sigma_{ik} = (i); \quad (N-1) \Sigma x_i^2 \geq \Sigma_{i \neq k} x_i x_k \iff (N-1) \Sigma x_i^2 - \Sigma x_i (N x - x_i) = (N x)^2 - \Sigma x_i^2 \iff N(\Sigma x_i^2) \geq (N x)^2
\]

which is Cauchy Schwarz.

**Lemma 2** With the notation as above for \( \rho, \sigma, c \); and the average correlation

\[
\rho^* = \frac{\Sigma_{i \neq k} \rho_{i,k}}{N (N-1)}:
\]

(1) If \( \sigma_i = \sigma_k \) for all \( i \neq k \), then \( \rho = \rho \backslash \rho_0^{\sigma^*} \)

(2) \( \rho \leq 1 \).

(3) If \( \rho = 1 \) then \( \rho_{i,k} = 1 \) for all \( i \neq k \)

**Proof:** (1) is immediate.

(2) \( \Sigma_{i \neq k} (\sigma_i - \sigma_k)^2 \geq 0 \iff 2N (N-1) \sigma^2 \geq 2 \Sigma_{i \neq k} \sigma_i \sigma_k \)

and using \( c_{ik} \leq \sigma_i \sigma_k \) it follows that \( \sigma^2 \geq c \)

(3) Suppose \( \rho = c / \sigma^2 = 1 \);

then

\[
\Sigma_{i \neq k} c_{ik} / N (N-1) = \Sigma \sigma_i^2 / N, \text{ or } \Sigma_{i \neq k} c_{ik} = (N-1) \Sigma \sigma_i^2 \leq \Sigma_{i \neq k} \sigma_i \sigma_k
\]

since \( c_{ik} \leq \sigma_i \sigma_k \). However, from lemma 1 we see that \( (N-1) \Sigma \sigma_i^2 \geq \Sigma_{i \neq k} \sigma_i \sigma_k \). Hence \( \Sigma \sigma_i^2 = \Sigma_{i \neq k} \sigma_i \sigma_k \) and by Cauchy Schwarz all \( \sigma_i \) are the same. By (1) \( \rho = \rho^* \).

Since each \( \rho_{i,k} \leq 1 \) and the \( \rho^* = 1 \), it follows that \( \rho_{i,k} = 1 \).

The following lemma provides a convenient way to predict the correlations of large aggregations based on estimates of global correlation from smaller aggregations.

**Lemma 3:** Write \( \rho (N) = \rho (\Sigma_{i \neq k} X_i, \Sigma_{i \neq k} Y_i) \)

For large \( N \), \( \rho (N) \) is approximated by the continuous function

\[
f(x) = 1 - e^{-\frac{1}{2} \rho (N) (x^2 - 1)}
\]

**Proof:** Solving \( \rho (N) = N \rho / [1 + (N-1) \rho] \) for the global correlation \( \rho \):

\[
\rho = \frac{\rho (N)}{N - \rho (N) (N-1)}
\]

Replace \( \rho (N) \) by \( f(x), x > 1 \). For \( 0 \leq \rho \leq 1 \) write:

\[
f(x) = \rho \left[ x - xf' (x) + f(x) \right].
\]

Differentiating both sides:

\[
f'(x) = r \left[ 1 - f(x) - xf'' (x) + f'(x) \right] \]

\[
f'(x) [1 + \rho (x-1)] = \rho [1 - f(x)]
\]

\[
f'(x) / [1 - f(x)] = -d \ln [1 - f(x)] = \rho / [1 + \rho (x-1)]
\]

\[
f(x) = 1 - e^{-\frac{1}{2} \rho (N) \rho' / [1 + \rho (u-1)]}
\]

Example, Crop Loss

Crop loss claims per US county per year are tabulated from 1980 – 2008 (data available at http://www.rff.org/events/event/data-climate-change-and-extreme-events). Restricting to counties without zero entries, a dataset of 1334 counties is obtained. For this dataset the average variance over all counties and the average covariance between pairs of counties can be computed. Their ratio is the global correlation, 0.103, as shown in Table 1. Random aggregation of disjoint pairs of size 20, 50, 100 and 200 counties are also constructed and correlations of the aggregates are computed. Iterating this process 2000 times, the correlations of disjoint randomly drawn aggregates are estimated by averaging over the 2000 iterations. Plugging these estimates into eqn (5) yields estimates of the global correlation, also shown in Table 1.

To illustrate the use of eqn (6), suppose the global correlation is estimated by averaging the correlations of 2000 samples of disjoint pairs of counties of size 20. The value from Table 1 is 0.120. Plugging this value of \( \rho \) into eqn (6), the curve \( f(x) \), approximating \( \rho (N) \) is plotted in Figure 1. The true values of \( \rho (N) \) computed with the true global correlation 0.103 are given for \( N = 20, 50, 100, 200 \). In this case, averaging the correlations of 2000 aggregations
of size 20 would give a reasonable estimate of the global correlation and of the correlations of larger aggregations.

Example: Uncertainty in Global Forest Carbon

There are $11.3 \times 10^9$ global hectares of biologically productive surface, of which approx $4 \times 10^9$ are forested. The terrestrial biosphere reservoir contains carbon in organic compounds in vegetation living biomass ($450$ to $650$ PgC, IPCC AR5 https://www.ipcc.ch/report/ar5/). Houghton et al (2009) [10] give $385 \sim 650$ GtC, stating that $70 \sim 90\%$ of that pool as forest. Using 80% gives a range of $360 \sim 520$ (IPCC) or $308 \sim 520$ (Houghton) GtC in Earth’s forests. The IPCC values give a forest carbon global density range of $90 \sim 130$ tC/ha. Assuming that $360$ and $520$ GtC are two independent samples from our uncertainty on the global forest carbon pool, we may ballpark this uncertainty as

$$\text{VAR (global forest carbon pool) } \sim \sqrt{2} (160)^2 [\text{GtC}]^2.$$  

$$\text{STD (global forest carbon pool) } = 113 \times 10^9 \ [\text{tC}].$$

Using (2):

$$113 \times 10^9 = \sigma 4\times 10^9 (4 \times 10^9)^{-1} + \rho)^{\frac{1}{2}} [\text{tC}].$$

$$28.3 = \sigma(2.5E-10 + \rho)^{\frac{1}{2}}. \quad (7)$$

where $\sigma$ is the root of the average variance of forest carbon in [t/ha], and $\rho = c / \sigma^2$. The challenge is to find values of $\sigma$ and $\rho$ that “harmonize” with uncertainty in forest carbon at the global level and the mean density of $90 \sim 130$ tC/ha.

If $\rho = 0$, then $\sigma = 1.8 \times 10^6$ tC. This would be an extremely fat tailed distribution that is not prime facie plausible. If $\rho = 1$, then the average uncertainty (standard deviation) of tC/ha would be 28.3. In itself, this value is not preposterous, but $\rho = 1$ is. In this case lemma 2.3 entails that the uncertainty of the carbon in any two hectares is perfectly correlated.

Weisbin et al (2013) [11] suggest $\sigma$ is in the order of 10% of the measured value up to 100 t/ha, linearly interpolated between 10% and 30% up to 150 tC. For the above global density range, that yields an estimate of $\sigma = 9 \sim 18$. Putting $\sigma = 9 \sim 18$ tC/ha in (4), we get $\rho = 2.5 \sim 9.9$, which is impossible.

Either the estimates of uncertainty at the global level (LHS eqn(7)) must come down or the uncertainty at the hectare scale ($\sigma$) must be larger than suggested in Weisbin et al [11], in order that the two can be combined with a plausible value of $\rho$ in eqn(7). If $\rho = 0.1$ then $\sigma = 89.5$ tC which is in the range of the average density but larger than expected on the basis of existing literature.

CONCLUSION

Correlations of random aggregations can be used to estimate global correlation. This quantity is important when trying to relate uncertainty at global scales to uncertainty at local scales. The IPCC AR5 estimates of uncertainty in global forest carbon must come down, or local estimates of uncertainty in carbon measurements per hectare must go up to achieve consistency.

REFERENCES
1. Hotelling H. Relations Between Two Sets of Variates. Biometrika. 1936; 28: 321-377.

2. Koch, Gary G. Intraclass correlation coefficient. In Samuel Kotz and Norman L. Johnson. Encyclopedia of Statistical Sciences. 4. New York: John Wiley & Sons. 1982; 213-217.

3. Kendall M, Stuart A. The Advanced Theory of Statistics Volume 2 Inference and Relationship. Charles Griffin And Company, London, UK. 1961.

4. Holmes RB. On random correlation matrices. SIAM Matrix Anal. 1991; 12.

5. Vu HV, Nguyen HH. Random matrices: law of the determinant. 2012.

6. Cooke RM. Markov and Entropy Properties of Tree and Vine-Dependent Variables. In Proceedings of the Section on Bayesian Statistical Science. American Statistical Association. 1997.

7. Bedford TJ, Cooke RM. Vines - A New Graphical Model for Dependent Random Variables. Annals of Statistics. 2002; 30: 1031-1068.

8. Joe H. Generating random correlation matrices based on partial correlations. Journal of Multivariate Analysis. 2006; 97: 2177-2189.

9. Kurowicka D, Joe H, Lewandowski D. Generating random correlation matrices based on vines And Extended Onion Method, Journal of Multivariate Analysis. 2009; 100: 1989-2001.

10. Houghton RA, Forrest H, Goetz SJ. Importance of biomass in the global carbon cycle, J of Geophysical Research. 2009; 114.

11. Weisbin CR, Lincoln W, Saatchi S. A Systems Engineering Approach to Estimating Uncertainty in Above-Ground Biomass (AGB) Derived from Remote-Sensing Data. Systems Engineering. 2013.