Research Article **Open Access**

Roger M. Cooke*, Sassan Saatchi, and Stephen Hagen

Global correlation and uncertainty accounting

DOI 10.1515/demo-2016-0009
Received May 23, 2016; accepted July 12, 2016

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.

Keywords: global correlation, forest carbon, uncertainty accounting.

MSC: 62H11, 62P12, 62P30.

1 Introduction

This note defines global correlation, studies its elementary properties and illustrates its use in global uncertainty accounting for crop loss and forest carbon. The rich literature on multivariate correlation can receive only passing mention. Conical correlation [6] concerns the maximal product moment correlation between linear combinations of two random vectors, interclass correlation [10] describes the correlations in grouped data. Multiple correlation and the correlation ratio [9] relate a single variable to a set of variables. Random correlation matrices [5] and the distribution of their determinants [4, 13] have sparked interest in the (scaled) determinant of the correlation matrix as a measure of multivariate association. Using Vines [1, 2, 8, 12] have made progress in understanding random determinants of correlation matrices. Micro correlations have attracted attention for their role in limiting the extent of securitization and risk sharing [11], and also for their role in amplifying tail dependence [3]. The problems discussed here involve up to 4 billion variables, and harmonizing uncertainty quantification at different scales of aggregation requires new techniques.

2 Methods

The correlation of random aggregates is used to estimate global correlation. All random variables are assumed to have a finite second moment. The following facts and definitions are used:

1) If X_1, X_2 are iid random variables with standard deviation σ, then

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

2) If $X_1, \ldots, X_N$ have average variance σ^2 and average covariance c, defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N} \mathrm{VAR}(X_i)}{N} \quad \text{and} \quad c = \frac{\sum_{i \neq k} \mathrm{COV}(X_i, X_k)}{N(N-1)},$$

***Corresponding Author: Roger M. Cooke:** Resources for the Future, E-mail: cooke@rff.org
Sassan Saatchi: Jet Propulsion Laboratory California Institute of Technology
Stephen Hagen: Applied GeoSolutions

© 2016 Roger M. Cooke et al., published by De Gruyter Open.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.

then

$$\mathrm{VAR}\left(\sum_{i=1}^{n} X_i\right) = N\sigma^2 + N(N-1)c \tag{1}$$

and, consequently, $c \geq -\sigma^2/(N-1)$.

3) Define $\underline{\rho} = c/\sigma^2$ as the *global correlation* of $X_1, \ldots, X_N$. Let $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$ have average variance σ^2 and average covariance c, both within and between components. That is

$$\sigma^2 = \frac{\sum_{i=1}^{N} \mathrm{VAR}(X_i)}{N} = \frac{\sum_{i=1}^{N} \mathrm{VAR}(Y_i)}{N}$$

and

$$c = \frac{\sum_{i\neq k} \mathrm{COV}(X_i, X_k)}{N(N-1)} = \frac{\sum_{i\neq k} \mathrm{COV}(Y_i, Y_k)}{N(N-1)} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{N} \mathrm{COV}(X_i, Y_k)}{N^2}.$$

Then from (1) one obtains

$$\rho\left(\sum_{i=1}^{N} X_i, \sum_{i=1}^{N} Y_i\right) = \frac{N^2 c}{N\sigma^2 + N(N-1)c} = \frac{N\underline{\rho}}{1 + (N-1)\underline{\rho}}. \tag{2}$$

The above correlation converges to 1 as $N \to \infty$, for any $\underline{\rho} > 0$. If $\underline{\rho} > 0$ and $N >> 1$, then

$$\mathrm{StDev}\left(\sum_{i=1}^{N} X_i\right) = \sigma N(N^{-1} + \underline{\rho}(N-1)/N)^{1/2} \sim \sigma N \underline{\rho}^{1/2}.$$

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

$$\mathrm{StDev}\left(\sum_{i=1}^{N} X_i\right) = \sigma N^{1/2}.$$

With independence, the uncertainty (standard deviation) of a sum of N random variables grows with $N^{1/2}$, but a small global correlation causes the growth to be linear in N. To appreciate this, let $\underline{\rho}$ 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 $N = 4E9$. With $\underline{\rho} = 0.001$, we have

$$\frac{\sigma N \underline{\rho}^{1/2}}{\sigma N^{1/2}} = 2000.$$

The difference between the cases $\underline{\rho} = 0$ and $\underline{\rho} = 0.001$ is huge. Recall the
Cauchy-Schwarz Inequality: for any $x, y \in \mathbb{R}^N$, we have

$$\left(\sum_{i=1}^{N} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right). \tag{3}$$

Equality in (3) holds if and only if $y_i = Ax_i$ for some $A \in \mathbb{R}$. With $y_i = 1$, we have

$$\left(\sum_{i=1}^{N} x_i\right)^2 \leq N \sum_{i=1}^{N} x_i^2, \tag{4}$$

with equality if and only if the x_i are constant. Equivalently, if $x = \sum_{i=1}^{N} x_i/N$ we have

$$N\sum_{i=1}^{N} x_i^2 \geq (Nx)^2 \tag{5}$$

or $x^2 \leq \sum_{i=1}^{N} x_i^2/N$ (a version of Jensen's inequality) with equality if and only if the x_i are constant.

3 Results and Discussion

Lemma 1. *For all* $x \in \mathbb{R}^N$, *we have* $(N-1)\sum_{i=1}^N x_i^2 \geq \sum_{i\neq k} x_i x_k$.

Proof. Put $x = \sum_{i=1}^N x_i/N$. Then $(N-1)\sum_{i=1}^N x_i^2 \geq \sum_{i\neq k} x_i x_k \Leftrightarrow (N-1)\sum_{i=1}^N x_i^2 \geq \sum_{i=1}^N x_i(Nx - x_i) = (Nx)^2 - \sum_{i=1}^N x_i^2 \Leftrightarrow N\sum_{i=1}^N x_i^2 \geq (Nx)^2$ which is (5). □

Lemma 2. *With the notation as above for* $\underline{\rho}$, σ, c; *with* $\sigma_i = \mathrm{VAR}(X_i)^{1/2}$, $c_{ik} = \mathrm{COV}(X_i, X_k)$, $\rho_{ik} = c_{ik}/(\sigma_i\sigma_k)$ *and the* average correlation *defined as*

$$\rho^\star = \frac{\sum_{i\neq k}\rho_{ik}}{N(N-1)},$$

we have:

(i) *If* $\sigma_i = \sigma_k$ *for all* $i \neq k$, *then* $\underline{\rho} = \rho^\star$;
(ii) $\underline{\rho} \leq 1$;
(iii) *If* $\underline{\rho} = 1$ *then* $\rho_{ik} = 1$ *for all* $i \neq k$.

Proof. (i) is immediate. (ii) $\sum_{i\neq k}(\sigma_i - \sigma_k)^2 \geq 0 \Leftrightarrow 2N(N-1)\sigma^2 \geq 2\sum_{i\neq k}\sigma_i\sigma_k$ and using $c_{ik} \leq \sigma_i\sigma_k$ it follows that $\sigma^2 \geq c$. (iii) Suppose $\underline{\rho} = c/\sigma^2 = 1$; then $\sum_{i\neq k} c_{ik}/(N(N-1)) = \sum_{i=1}^N \sigma_i^2/N$ or $\sum_{i\neq k} c_{ik} = (N-1)\sum_{i=1}^N \sigma_i^2 \leq \sum_{i\neq k}\sigma_i\sigma_k$ since $c_{ik} \leq \sigma_i\sigma_k$. However, from Lemma 1 we see that $(N-1)\sum_{i=1}^N \sigma_i^2 \geq \sum_{i\neq k}\sigma_i\sigma_k$. Hence $(N-1)\sum_{i=1}^N \sigma_i^2 = \sum_{i\neq k}\sigma_i\sigma_k$ or $N\sum_{i=1}^N \sigma_i^2 = (\sum_{i=1}^N \sigma_i)^2$. By the Cauchy-Schwarz inequality (see (4)) all the σ_i are the same. By (i) $\underline{\rho} = \rho^\star = 1$. Since each $\rho_{ik} \leq 1$ and $\rho^\star = 1$, it follows that $\rho_{ik} = 1$. □

Writing $\rho_N = \rho\left(\sum_{i=1}^N X_i, \sum_{i=1}^N Y_i\right)$ we construct a continuous version of ρ_N as follows. Solve (2) for the global correlation $\underline{\rho}$:

$$\underline{\rho} = \frac{\rho_N}{N - (N-1)\rho_N}. \tag{6}$$

Replace ρ_N by $f(x)$, $x > 1$. For $0 \leq \underline{\rho} \leq 1$ write:

$$f(x) = \underline{\rho}[x - xf(x) + f(x)]. \tag{7}$$

Differentiating both sides of (7):

$$f'(x) = \underline{\rho}[1 - f(x) - xf'(x) + f'(x)],$$

$$f'(x)[1 + \underline{\rho}(x-1)] = \underline{\rho}[1 - f(x)],$$

$$\frac{f'(x)}{1-f(x)} = -\frac{d[\ln(1-f(x))]}{dx} = \frac{\underline{\rho}}{1+\underline{\rho}(x-1)},$$

$$f(x) = 1 - \exp\left(-\int_{1<u\leq x} \frac{\underline{\rho}}{1+\underline{\rho}(u-1)}du\right). \tag{8}$$

Equation (8) provides a graphical representation of the relation between ρ_N and $\underline{\rho}$ (see Figure 1).

4 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 1, 10, 20, 50, 100 and 200 counties are also constructed and correlations of the aggregates are computed. Iterating this process 2000 times, the correlation of disjoint randomly drawn aggregates are estimated by averaging over the 2000 iterations. Plugging these estimates into (6) yields estimates of the global correlation, also shown in Table 1.

Table 1: US crop loss insurance claims from 1330 counties without null entries, 1980 - 2008.

Estimates of Global Correlation from Aggregate Correlation, 2000 iterations						
Aggregation size	1	10	20	50	100	200
Average correlation	1.98E-01	5.89E-01	7.30E-01	8.73E-01	9.31E-01	9.64E-01
STDev of correlations	2.64E-01	1.71E-01	1.11E-01	4.92E-02	2.47E-02	1.33E-02
Global correlation estimate	1.97E-01	1.30E-01	1.20E-01	1.18E-01	1.16E-01	1.16E-01
Average Variance	6.33E+12					
Average Covariance	6.49E+11					
Global Correlation	1.03E-01					

To illustrate the use of (8), 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 $\underline{\rho}$ into (8), the curve $f(x)$ approximating ρ_N is plotted in Figure 1. The true values of ρ_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.

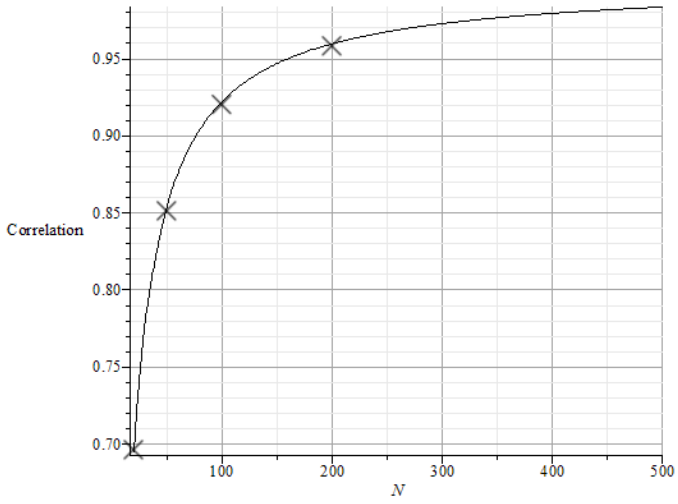

Figure 1: Plot of $f(x)$ (see (8)) using $\underline{\rho}$ estimated from 2000 aggregations of size 20, and true values of ρ_N, $N = 20, 50, 100, 200$, computed with the actual global correlation.

5 Example: Uncertainty in global forest carbon

There are 11.3E09 global hectares of biologically productive surface, of which approx 4E09 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/). [7] gives 385 - 650 GtC, stating that $70 \sim 90\%$ of that pool as forest. Using 80% gives a range of $360 \sim 520$ (IPCC) or $308 \sim 520$ [7] 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

$$\text{VAR}(\textit{global forest carbon pool}) \sim 1/2(160)^2\ [\text{GtC}]^2.$$

$$\text{StDev}(\textit{global forest carbon pool}) = 113\text{E}09\ [\text{tC}].$$

Using (1):

$$113\ \text{E}09 = \sigma 4\text{E}09((4\text{E}09)^{-1} + \underline{\rho})^{1/2}[\text{tC}],$$

$$28.25 = \sigma(2.5\text{E}{-10} + \underline{\rho})^{1/2}, \tag{9}$$

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

If $\underline{\rho} = 0$, then $\sigma = 1.8\text{E}06$ tC. This would be an extremely fat tailed distribution that is not prima facie plausible. If $\underline{\rho} = 1$, then the average uncertainty (standard deviation) of tC/ha would be 28.3. In itself, this value is not preposterous, but $\underline{\rho} = 1$ is. In this case Lemma 2(iii) entails that the uncertainty of the carbon in any two hectares is perfectly correlated.

[14, Table II] suggest σ is in the order of 10% of the measured value up to 100 tC/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 (9), we get $\underline{\rho} = 2.5 \sim 9.9$, which is impossible.

Either the estimates of uncertainty at the global level (LHS of (9)) must come down or the uncertainty at the hectare scale (σ) must be larger than suggested in [14], in order that the two can be combined with a plausible value of $\underline{\rho}$ in (9). If $\underline{\rho} = 0.1$ then $\sigma = 89.5$tC which is in the range of the average density but larger than expected on the basis of existing literature.

6 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. Statistical properties of estimators of global correlation remain to be explored, and more inequalities between global and average correlation can probably be found.

References

[1] Bedford, T. and R. M. Cooke (2002). Vines – a new graphical model for dependent random variables. *Ann. Statist. 30*(4), 1031–1068.

[2] Cooke, R. M. (1997). Markov and entropy properties of tree and vine-dependent variables. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria VA.

[3] Cooke, R. M., C. Kousky, and H. Joe (2011). Micro correlations and tail dependence. In *Dependence modeling*, pp. 89–112. World Sci. Publ., Hackensack, NJ.

[4] Hanea, A.M. and G.F. Nane, (2016). The Asymptotic Distribution of the Determinant of a Random Correlation Matrix. Preprint available at http://arxiv.org/abs/1309.7268.
[5] Holmes, R. B. (1991). On random correlation matrices. *SIAM J. Matrix Anal. Appl. 12*(2), 239–272.
[6] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika 28*(3/4), 321–377.
[7] Houghton, R. A., F. Hall, and S. J. Goetz (2009). Importance of biomass in the global carbon cycle. *J. Geophys. Res. Biogeosci. 114*(G2).
[8] Joe, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivariate Anal. 97*(10), 2177–2189.
[9] Kendall, M. G. and A. Stuart (1967). *The Advanced Theory of Statistics. Vol. 2: Inference and Relationship.* Second edition. Hafner Publishing Co., New York.
[10] Koch, G. G. (2006). Intraclass correlation coefficient. In *Encyclopedia of Statistical Sciences 6*. John Wiley & Sons, New York.
[11] Kousky, C. and R. M. Cooke (2011). The limits of securitization: micro-correlations, fat tails and tail dependence. In K. Boecker (Ed.), *Re-Thinking Risk Measurement and Reporting, Uncertainty, Bayesian Analysis and Expert Judgement*, pp. 295–330. Risk Books, London.
[12] Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal. 100*(9), 1989–2001.
[13] Nguyen, H. H. and V. Vu (2014). Random matrices: law of the determinant. *Ann. Probab. 42*(1), 146–167.
[14] Weisbin, C. R., W. Lincoln, and S. Saatchi (2014). A systems engineering approach to estimating uncertainty in above-ground biomass (agb) derived from remote-sensing data. *Syst. Engin. 17*(3), 361–373.