Probabilistic aggregation of uncertain geologic resources

Gordon M. Kaufman[1] • Ricardo A. Olea[2] • Ray Faith[3] • Madalyn S. Blondes[4] •

Abstract Commodities such as oil and gas occur in isolated reservoirs or accumulations, more generically called basic units here. To understand a study area's economic potential and to craft plans for exploration and development, resource analysts often aggregate (sum, accumulate) basic unit magnitudes in distinct spatial subsets of the study area and then appraise the total area's potential by summing these intermediate sums. In a probabilistic approach, magnitudes are modeled as random variables. Some have asked, "Do different methods of partitioning basic units into subsets lead to different probability distributions for the sum of all basic unit magnitudes?" Any method of aggregation of basic unit magnitudes which obeys the rules of probability leads to the same probability distribution of the sum of all unit magnitudes as that computed by direct summation of all basic unit magnitudes. A Monte Carlo simulation of a synthetic example in which the magnitude of resource in each unit is marginally lognormal and pairwise correlations among basic unit magnitudes are specified illustrates key features of probabilistic aggregation. The joint distribution of certain pairs of aggregates are closely approximated by a bivariate lognormal distribution.

[1] Gordon Kaufman
Sloan School of Management
Massachusetts Institute of Technology
e-mail:gkaufman@mit.edu

[2]Ricardo A. Olea
Eastern Energy Resources Science Center
U.S. Geological Survey, Reston, VA, USA
e-mail: rolea@usgs.gov

[3]Ray Faith
Sloan School of Management
Massachusetts Institute of Technology
e-mail: faith@mit.edu

[4]Madalyn Blondes
Eastern Energy Resources Science Center
U.S. Geological Survey, Reston,, VA, USA
e-mail: mblondes@usgs.gov

1 Introduction

Basic mineral commodities such as oil, gas, copper, silver and gold occur as isolated reservoirs (deposits, accumulations). A principal goal of studies that treat mineral magnitudes in individual accumulations as uncertain quantities—random variables (rvs)— is to provide probabilistic projections of magnitude totals of selected subsets of the collection of all accumulations in a (spatial) assessment frame as well as of the total of all accumulation magnitudes in the frame.

Mineral resources of the type studied here are geographically discontinuous and disseminated over a study area, which can be as large as a country or a continent. These accumulations are three-dimensional objects whose spatial distribution often appears as a geographic map. Geologists customarily define a "basic unit" to be, say, a copper deposit or a petroleum reservoir, field or play, and assign a probability distribution to the magnitude of resources in each unit leading to as many marginal probability distributions as basic units.

Direct probabilistic aggregation of a collection of basic unit magnitudes into a sum of all unit magnitudes requires specification of a *joint* probability distribution incorporating probabilistic dependencies among basic unit magnitudes, a daunting exercise when the number of units is large. Many resource assessment studies do specify a joint probability distribution of basic unit magnitudes (Carter and Morales 1998; Schuenemeyer 2005; Delfiner and Barrier 2008; Pike 2008; Schuenemeyer and Gautier 2010; Van Elk and Gupta 2010; Blondes et al. 2013a and 2013c; U.S. Geological Survey Carbon Dioxide Storage Resources Assessment Team 2013). Crovelli and Balay (1991) characterize dependencies among basic units and between aggregates in terms of covariances and correlations.

Methodological issues that arise in the course of probabilistic aggregation are the subject of this contribution. In Sect. 2 two proofs show that if the rules of probability are obeyed distinct partial intermediate aggregations of basic unit magnitudes lead to the same distribution of the sum of all basic unit magnitudes. Said differently, multiple levels of aggregation (multiple stage aggregation) lead to the same probability distribution for the sum of all accumulation magnitudes as direct summation of all accumulation magnitudes (single stage aggregation). Sect. 3 outlines properties of the data generating process used in Sect. 4 numerical examples. As a set of basic units is aggregated into a smaller number of larger sets do (positive) pairwise correlations between sums of basic unit magnitudes increase, decrease or stay the same? Is there an ordinal ordering of pairwise correlations among these sums as the number of elements in them increases? Sect. 5 addresses these questions, presents easy to compute bounds on allowable background correlations and establishes useful inequalities governing differences between pairwise correlations between basic unit magnitudes and pairwise correlations among aggregates of them. Remarks about practical aspects of aggregation and elicitation of geologic judgments appear in Sect. 6.

2 Aggregation

For many geological resources, such as oil and gas, there is a natural hierarchy of aggregation levels: individual accumulation magnitudes in an oil and gas field, the sum of individual magnitudes in a play, the sum of magnitudes in the collection of plays in a petroleum basin and in turn, a regional basin aggregate.

A first principle is:

If the laws of probability laws are obeyed the probability distribution of the sum of all individual accumulation magnitudes in a sample frame is

the same as the probability distribution for this sum computed by use of an aggregation scheme—no matter how one chooses to aggregate.

Define aggregation as follows: assume that choice of labelling of magnitudes is non-informative. Partition a set of *N* uncertain magnitudes $\{X_1,....,X_N\}$ into K mutually exclusive and collectively exhaustive subsets $A_1,..,A_K$. Define S_k to be the sum of elements in $A_K, k = 1,...,K$. Then $\{S_1,...,S_K\}$ is an *aggregate* of $\{X_1,...,X_N\}$.

Assertion 1: The cumulative distribution function of the sum $S = X_1 + ...X_N$ of *N* uncertain accumulation magnitudes (*rvs*) is identical to that of the sum $S_1 + ... + S_K$ of aggregates $S_1,...,S_K$.

The following simple proofs extend to successive levels of aggregation of $S_1,...,S_K$.

First Proof: Each possible realization $x_1,...,x_N$ of $X_1,...,X_N$ is a set of *N* real numbers, each in $(-\infty,\infty)$. Use parentheses to partition $x_1 + ... + x_N$ as $(x_1 + ... + x_{i_1}) + (x_{i_1+1} + ... + x_{i_2}) + ... + (x_{i_{K-1}+1} + ... + x_N)$. Sum numbers within each pair of parentheses and set $s_1 = x_1 + ... + x_{i_1}, s_2 = x_{i_1+1} + ... + x_{i_2}, ..., s_K = x_{i_{K-1}+1} + ... + x_N$. Numbers $x_1,...,x_N$ obey the associative law of arithmetic so $s_1 + s_2 + ... + s_K = x_1 + x_2 + ... + x_N$ for any such partition of $\{x_1,...,x_N\}$. This obtains for each possible realization $x_1,...,x_N$ of $X_1,...,X_N$ and all possible partitions of $\{x_1,...,x_N\}$ so the associative law of arithmetic applies to $X_1,...,X_N$ as well.

Second Proof: Suppose that the range of each $X_1,...,X_N$ is $(-\infty,\infty)$ and that the $(N\times 1)$ array of uncertain quantities $\mathbf{X}_N=(X_1,...,X_N)$ possesses a continuous probability distribution (density) $Prob\{\mathbf{X}_N\in d\mathbf{X}\}=f(\mathbf{X})d\mathbf{X}$ with respect to Lebesgue measure on $(-\infty,\infty)^N$. Compare the cumulative distribution function $F(s)\equiv Prob\{X_1+...+X_N\le s\}$ of $X_1+...+X_N$ with the cumulative distribution function $G_K(s)\equiv Prob\{S_1+...+S_K\le s\}$ of $S_1+...+S_K$. By construction $S_1=(X_1+...+X_{n_1}),S_2=(X_{n_1+1}+...+X_{n_2}),...,S_K=(X_{n_{K-1}+1}+...+X_N)$ so on applying the associative law of arithmetic to *rvs* $S_1,...,S_K$, $G_K(s)=F(s)$ for all $s\in(-\infty,\infty)$.

Assertion 1 obtains irrespective of the structure of dependencies assigned to $X_1,...,X_N$. Distinct aggregation schemes, each of which obey the rules of probability, lead to identical moments of all order for the sum of all basic unit magnitudes. (Assertion 2 below).

Blondes et al. (2013) make three assertions: first, that the probability distribution of an aggregated sum using multiple stages of correlation matrices is strongly dependent on the number of aggregation stages, the size of the individual groups, and the size of the total aggregation. Second, multiple stage aggregation will, if correlation coefficients are positive, narrow aggregate distributions. Third, the choice of partition of units into groups can have a larger impact on the distribution of the sum of all unit magnitudes than choices of correlation coefficient selected by experts. All or any of these assertions may obtain if geologists' probability judgments do *not* adhere to laws of probability. In contrast, Assertion 1 says that any aggregation scheme obeying the rules of probability leads to the same probability distribution as that for the sum of all basic unit magnitudes.

2.1 Moments and Aggregates

One tactic for simplifying the task of assessing properties of a large collection of basic units is to assume that the joint distribution of magnitudes $X_1,...,X_N$ is a parametric distribution indexed by a mean vector, a variance matrix and possibly a small number of additional parameters. Even so, the task of specifying parameters can be daunting. If *N* is large, the covariance matrix of $X_1,...,X_N$ possesses an intimidatingly large number of parameters. Aggregation of $X_1,...,X_N$ into $K << N$ subsets helps in principle: an analyst must then assess or estimate from available data $K(K+1)/2 << N(N+1)/2$ variances and co-variances of sums.

Turn next to an important property of variance matrices induced by aggregation. Define the vector rv $(N \times 1)\,\mathbf{X}_N = (X_1,...,X_N)^t$ and the variance matrix of $\mathbf{X}_N$ to be

$$Var(\mathbf{X}_N) = \mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & \vdots & \vdots \\ \vdots & \vdots & \ddots & v_{N-1,N} \\ v_{N1} & \cdots & v_{N,N-1} & v_{NN} \end{bmatrix}. \qquad (1)$$

Let $\mathbf{S}_K = (S_1,...,S_K)^t$ be a $(K \times 1)$ vector of aggregates and define $(K \times N)\mathbf{A}$ to be a matrix that maps elements of $\mathbf{X}_N$ into $K < N$ distinct sums $S_1,...,S_K$ as defined in Sec. 1. Place N_k 1s in the k^{th} row $\mathbf{a}_k$ of matrix $\mathbf{A}$ at labels of elements in A_k and 0s elsewhere so that $\mathbf{a}_k\mathbf{X}_N = S_k$, the sum of magnitudes in the subset A_k of elements of $\mathbf{X}_N$. Do this for all *K* subsets. Then

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_K \end{bmatrix} \tag{2}$$

maps $\mathbf{X}_N \rightarrow \mathbf{S}_K = (S_1,..,S_K)^t$. That is, $\mathbf{AX}_N = \mathbf{S}_K$ and $Var(\mathbf{S}_k) = Var(\mathbf{AX}_N) = \mathbf{A}Var(\mathbf{X}_N)\mathbf{A}^t$. In turn, on defining $(1\times K)\mathbf{1}_K$ to be a vector of 1s $Var(S) = \mathbf{1}_K^t\mathbf{AXA}^t\mathbf{1}_K$. The punchline is that $\mathbf{1}_K^t\mathbf{A} = \mathbf{1}_N$, a vector of *N* 1s so that $\mathbf{1}_K^t\mathbf{AXA}^t\mathbf{1}_K = \mathbf{1}_N^tVar(\mathbf{X}_N)\mathbf{1}_N = Var(S)$. This establishes by computation

> **Assertion 2**: The variance of the sum of elements of $\mathbf{X}_N$ equals the variance of the sum of aggregates of elements of $\mathbf{X}_N$ for all possible partitions of elements of $\mathbf{X}_N$ into non-null subsets.

Assertion 2 is, of course, a direct consequence of Assertion 1.

2.2 Multiple Stage Aggregation

Matrices of type displayed in Eq. (2) yield a compact representation of means and variances of basic unit sums induced by multiple levels of aggregation. For any fixed ordering of elements of $\mathbf{X}_N = (X_1,...,X_N)^t$, the matrix in Eq. (2) maps $\mathbf{X}_N$ into a vector of aggregates $\mathbf{S}_K = (S_1,...,S_K)^t$. Call $\mathbf{A}$ as in Eq. (2) $\mathbf{A}^{(1)}$ so that $\mathbf{A}^{(1)}\mathbf{X}_N = \mathbf{S}_K$.

Use a new version of Eq. (2) to aggregate the *K* elements of $\mathbf{S}_K$ into $M < K$ aggregates. Partition sums $\{S_1,...,S_K\}$ into M subsets $B_1,..,B_M$ with $M_1,...,M_M$ members respectively. Place M_m 1s in the m^{th} row of a new matrix at labels of elements of B_k and 0s elsewhere in that row. Then $\mathbf{a}_m^{(2)}\mathbf{S}_K = W_m$, the sum of magnitudes in the subset B_k of elements of $\mathbf{S}_K$. Do this for all *K* subsets. The $(M\times K)$ matrix

$$\mathbf{A}^{(2)} = \begin{bmatrix} \mathbf{a}_1^{(2)} \\ \mathbf{a}_2^{(2)} \\ \vdots \\ \mathbf{a}_K^{(2)} \end{bmatrix} \tag{3}$$

maps the K elements of $\mathbf{S}_K$ into $M < K$ aggregates $(1 \times M)\ \mathbf{W}_M = \mathbf{A}^{(2)}\mathbf{S}_K = \mathbf{A}^{(2)} \times \mathbf{A}^{(1)}\mathbf{X}_N$.

This leads to:

Assertion 3: If aggregation to a sum of all basic unit magnitudes is done in p stages, $\mathbf{A}^{(p)} \times ... \times \mathbf{A}^{(2)} \times \mathbf{A}^{(1)}$ is an $(N \times 1)$ row vector of $1s$ and

$$S = \mathbf{A}^{(p)} \times ... \times \mathbf{A}^{(2)} \times \mathbf{A}^{(1)}\mathbf{X}_N . \tag{4}$$

3 Dependencies

How best to appraise probabilistic dependencies among basic mineral resource units is a recurring issue—from the first large scale exercise in subjective geological assessment of basic mineral resources (Miller et al. 1975) to recent attempts. Authors of the Circum-Arctic study (Schuenemeyer and Gautier 2010) make it clear that probabilistic projections of oil and gas in this very large region are sensitive to variations in co-variabilities of basic unit magnitudes. They point out that when 48 Circum-Arctic assessment units are aggregated 90% uncertainty intervals for recoverable gas range from 1,471 TCF, to 2,009 TCF, to 3,515 TCF for assumptions of independence, assessor specified dependencies (correlations), and functional dependence of all units (Pearson correlation coefficient 1.0) respectively. Decision makers who rely on assessment results need accurate interval estimates. Too broad an interval provides little information; a too narrow interval gives a false sense of precision.

To keep the assessment task within bounds geologists often limit appraisal of dependencies to pairwise correlations among $X_1,...,X_N$ or among aggregates of them. In most realistic geological assessment exercises, pairwise correlations range from close to zero to close to 1.0 with large subsets of correlations in between. Several USGS studies (Collett 2008; Klett and Gautier 2009) state that zero pairwise correlation implies probabilistic independence of a pair of uncertain quantities and, at the opposite extreme, claim that assignment of correlation 1.0—often mislabeled as "perfect correlation"— allows computation of fractiles of a sum of all basic unit magnitudes by addition of basic unit fractiles. Neither statement is true in general.

3.1 Data Generating Process

In what follows, the probability law governing $\mathbf{X}_N$ is multivariate lognormal so the joint distribution of elements of the vector $\ln \mathbf{X}_N = (\ln X_1,...,\ln X_N)^t \equiv \mathbf{Y}_N$ is fully specified by assignment of a mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ to natural logarithms of elements of the basic magnitude vector $\mathbf{X}_N$:

$$Var(\mathbf{Y}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \sigma_{N-1,N} \\ \sigma_{N1} & \cdots & \sigma_{N,N-1} & \sigma_{NN} \end{bmatrix}. \quad (5)$$

The correlation matrix associated with $\boldsymbol{\Sigma}$ is

$$. \quad Corr(\mathbf{Y}) = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1N} \\ r_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & r_{N-1,N} \\ r_{N1} & \cdots & r_{N,N-1} & 1 \end{bmatrix}, \; r_{ij} = \sigma_{ij} / \sigma_{ii}\sigma_{jj}. \quad (6)$$

3.2 Basic Unit Magnitude Correlation Structure

The correlation structure of $\mathbf{X}_N = (X_1,...,X_N)^t$ and that of $\ln \mathbf{X}_N = (\ln X_1,...,\ln X_N)^t = \mathbf{Y}_N$ demand attention. First, the dispersion of a sum of lognormal *rvs* is a function of sums of pairwise co-variances among them and small variations in co-variances can lead to large differences in dispersion of this sum. Second, while in theory personal probability judgments about basic unit magnitudes do not depend on whether elicitation in done in units of magnitude or units of logarithms of magnitude, in practice distinct choices of scale and function often lead to distinct probability judgments about unit magnitudes even when they should not.

If $\mathbf{X} = (X_1, X_2)^t$ is lognormal (is $LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{7}$$

the mean of $\mathbf{X}$ is

$$E(\mathbf{X}) = \begin{pmatrix} \exp\{\mu_1 + \frac{1}{2}\sigma_{11}\} \\ \exp\{\mu_2 + \frac{1}{2}\sigma_{22}\} \end{pmatrix} \equiv \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \quad . \tag{8}$$

With v_{11}, v_{22} and v_{12} as in Eq. (1),

$$Var(X_1) = M_1^2(e^{\sigma_{11}} - 1), \;\; Var(X_2) = M_2^2(e^{\sigma_{22}} - 1), \tag{9}$$

and the covariance of X_1 and X_2 is

$$v_{12} = M_1 M_2 (e^{\sigma_{12}} - 1) \; . \tag{10}$$

The correlation coefficient of X_1 and X_2 is

$$Corr(X_1, X_2) = v_{12} / \sqrt{v_{11} v_{22}} = (e^{\sigma_{12}} - 1) / \sqrt{e^{\sigma_{11}} - 1} \times \sqrt{e^{\sigma_{22}} - 1} \ . \tag{11}$$

Modest variations in variances of logarithms of basic unit magnitudes can induce large variations in $Cov(X_1, X_2)$ and $Corr(X_1, X_2)$. However, when σ_{ii} and σ_{jj} are small

$$Corr(X_i, X_j) = r_{ij} \times (1 + O(\sigma_{ii} \sigma_{jj})) . \tag{12}$$

3.3 Covariance Structure of Aggregates

In the numerical study described below pairwise correlations among basic unit magnitudes are restricted to be positive. Consider 12 basic units partitioned into two subsets (clusters), one with 7 units and the other with 5 units. In Table 1 within cluster pair-wise correlations of 0.7 and 0.6 are assigned to green and blue clusters respectively and 0.4 to between cluster pairwise correlations.

(Table 1 Here)

The special structure of diagonal green and blue blocks in Table 1 allow calculation of simple formulae for variances of sums $S_1 = X_1 + ... + X_7$ and $S_2 = X_8 + ... X_{12}$. The green sub-matrix in Table 1 is a version of an intra-class correlation matrix. With $\mathbf{X}_1 = (X_1, ..., X_n)^t$, $\mathbf{X}_2 = (X_{n+1}, ..., X_N)^t$, within cluster correlations $\theta_1 = Corr(X_i, X_j)$ $i, j = 1, ..., n$ and i ≠ j, $\theta_2 = Corr(X_i, X_j)$ $i, j = n+1, ..., N$ and i ≠ j, between cluster correlations ρ, $(n \times 1)\mathbf{1}_n = (1, ..., 1)^t$ and $\mathbf{D}_a^{1/2} = diag(\sqrt{v_{11}}, ..., \sqrt{v_{nn}})$,

$$Var(\mathbf{X}_1) = \mathbf{D}_a^{1/2} [(1 - \theta_1)\mathbf{I}_n + \theta_1 \mathbf{1}_n \mathbf{1}_n^t] \mathbf{D}_a^{1/2} . \tag{13}$$

Using Eq. (13), $n = 7$ and $N = 12$

$$Var(S_1) = (1-\theta_1)(v_{11} + ... + v_{nn}) + \theta_1(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2, \quad (14)$$

$$Var(S_2) = (1-\theta_2)(v_{n+1,n+1} + ... + v_{NN}) + \theta_2(\sqrt{v_{n+1,n+1}} + ... + \sqrt{v_{NN}})^2 \quad (15)$$

and

. $$Cov(S_1, S_2) = \rho(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})(\sqrt{v_{n+1,n+1}} + ... + \sqrt{v_{N,N}}) \quad (16)$$

Standard deviations of basic unit magnitudes in Table 5 along with correlations as in Table 1 determine the variance matrix $Var(\mathbf{X}_N)$.

(Table 2 Here)

Standard deviations of $S_1 = X_1 + ... + X_7$ and $S_2 = X_8 + ... + X_{12}$ are 184 and 150 respectively, $Cov(S_1, S_2) =$ 179 and $Corr(S_1, S_2) = 0.562$. The variance of the sum of all 12 basic unit magnitudes is 87,381 and its standard deviation is 296.

Variance and correlation matrices for basic units aggregated into four clusters{1,2,3,4}, {5,6,7}, {8,9,10}, {11,12} appear in Table 3.

(Table 3 Here)

Table 4 displays variance correlation matrices for Table 4 clusters aggregated into two larger clusters {1,2,3,4,5,6,7} and {8,9,10,11,12}.

(Table 4 Here)

In Tables 3 and 4 correlations of aggregated sums are larger than the common pairwise correlation 0.4 assigned to pairs of individual basic unit magnitudes, each member of a pair in a distinct cluster.

4. Simulation

Properties of aggregates of basic unit magnitudes possessing marginal distributions shown in Table 5 and correlation structure as in Table 1 are computed by Monte Carlo simulation. A key assumption is that basic unit magnitudes are jointly lognormal.

Figure 1 displays boxplots of Table 5 marginal distributions:

(Table 5 Here)

(Figure 1 Here)

4.1 Numerical Aggregation

Figures 2 and 3 are overlays of the empirical cumulative distribution function of the sum of all twelve magnitudes and a fit of a lognormal distribution to this sum. They are virtually indistinguishable. This feature is possibly an artifact of the small variation in the range of σ^2 (0.202 to 0.631) and in the range of μ (3.004 to 4.200) among the twelve basic unit distributions in Table 5.

(Figure 2 Here)

(Figure 3 Here)

A Q-Q plot provides a finer pictorial resolution of right tail behavior. Figure 4 shows a lognormal fit to be surprisingly good out to the $15/100{,}000^{th} = 0.99985^{th}$ fractile beyond which right tail deviations are visible.

(Figure 4 Here)

Figures 5(a), 5(b) and 5(c) compare empirical marginal distributions of sums of magnitudes in clusters{1,2,3,4,5,6}and in {7,8,9,10,11,12}and of total magnitude with lognormal density fits using 100,000 Monte Carloed values.

(Figures 5(a), 5(b), 5(c) Here)

Figure 6(a) is a scatterplot of sums of magnitudes in clusters {1,2,3,4,5,6} and {7,8,9,10,11,12}. Figure 6(b) is a scatterplot of logarithms of sums of magnitudes in these two clusters.

(Figures 6(a), 6(b) Here)

4.2 Approximate Lognormality

The shape of the scatterplots in Fig. 6 suggests that a bivariate lognormal distribution may be a reasonable fit to the joint distribution of the logarithms of cluster sums. The fit to un-normalized sums is quite good for the particular set of correlated lognormal *rvs* used here. Figs. 7(a), 7(b) and 7(c) are Q-Q plots of the empirical distribution of the sum of magnitudes in cluster {8,9,10,11,12} conditioned on the sum of magnitudes in cluster {1,2,3,4,5,6,7} at slices taken at the median and ± 1 standard deviations from the median.

(Figures 7(a), 7(b), 7(c) Here)

5 Aggregation and Correlation

Do pairwise correlations between sums of basic unit magnitudes increase, decrease or stay the same as basic units are aggregated into smaller numbers of larger and larger sets? Is there an ordinal ordering of pairwise correlations among these sums as the number of elements in them increases? Answers to both questions are "No" in general. However, variance matrices structured as in the USGS Circum-Arctic study and in Table 1 lead to useful inequalities between pairwise correlations among individual unit magnitudes and correlations between sums of magnitudes. Partition the set of all basic units into two distinct subsets (clusters) A_1 and A_2 chosen so that the magnitude of any unit in A_1 and that of any unit in A_2 possess identical pairwise correlation. Sect. 5.2 provides a proof that, for positive background correlations, the pairwise correlation between the sum of unit magnitudes in A_1 and the sum of unit magnitudes in A_2 is uniformly larger than the common (background) correlation assigned to two individual units in distinct clusters. Sect. 5.1 sets the stage with presentation of properties of Shür complements used to show that, as the number of elements in A_1 and the number of elements in A_2 increase in accord with a uniform asymptotic regime described in Sect. 5.3, the pairwise correlation between A_1 and A_2 sums approaches a limit proscribed by a function of weighted averages of within cluster correlations.

5.1 Shür Complements

Consider the $(N \times N)$ variance matrix of $\mathbf{X}_1 = (X_1 + ... + X_n)^t$ and $\mathbf{X}_2 = (X_{n+1} + ... + X_N)^t$:

$$Var(\mathbf{X}) = Var\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \equiv \mathbf{V}, \quad (n \times n)\mathbf{V}_{11} \ . \tag{17}$$

Here elements of $\mathbf{X}_1$ ares interpretable as magnitudes of a cluster of geologically similar units for which geologists provide sufficient information to pin down numeric values for components of $\mathbf{V}_{11}$. Interpret $\mathbf{X}_2$ and $\mathbf{V}_{22}$ similarly. According to Schuenemeyer and Gautier (2010) correlations between two basic unit magnitudes lying in distinct clusters are not easy to pin down and the number of them their study is large. To limit complexity they assume that almost all pairwise correlations between two units in distinct clusters share a common value and call each such correlation "background correlation". Table 1 is a simple example in which pairwise correlations between basic unit magnitudes within each of two distinct clusters share a common value. Assertion 4 below documents how allowable values of background correlation depend on variance matrices assigned to clusters.

Necessary conditions for $\mathbf{V}$ as in Eq. (17) to be positive definite are first that $\mathbf{V}_{11}$ is PDS $(\mathbf{V}_{11} > 0)$ and second, that the Schür complement

$$\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} > \mathbf{0}. \tag{18}$$

Factor $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$ so that $\mathbf{A}_{11} = \mathbf{D}_1^{-1/2}\mathbf{V}_{11}\mathbf{D}_1^{-1/2}, \mathbf{D}_1^{1/2} = diag(\sqrt{v_{11}}, ..., \sqrt{v_{nn}})$ and $\mathbf{A}_{22} = \mathbf{D}_2^{-1/2}\mathbf{V}_{22}\mathbf{D}_2^{-1/2}, \mathbf{D}_2^{1/2} = diag(\sqrt{v_{n+1,n+1}}, ..., \sqrt{v_{NN}})$. Then $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are correlation matrices generated by $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$. The necessary condition Eq. (18) is equivalent to

$$\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} > \mathbf{0}\ ,\ \mathbf{A}_{12} = \mathbf{D}_1^{-/12}\mathbf{V}_{12}\mathbf{D}_2^{-/12}. \tag{19}$$

A version of the following assertion appears in Kaufman (2016) along with tighter but more recondite inequalities for patterned variance matrices.

Assertion 4:

(1) The sum of elements of the inverse of any PDS correlation matrix is strictly greater than one.

(2) Consider the correlation matrix $\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ associated with $\mathbf{V}$ as in (17) when $(n \times N - n)\mathbf{A}_{12} = \mathbf{1}_n \mathbf{1}_m^t \times \rho$ and $\mathbf{A}_{11}, \mathbf{A}_{22} > \mathbf{0}$. Define g_i to be the sum of elements of $\mathbf{A}_{ii}^{-1}$. Then $\mathbf{V}$ is PDS if and only if

$$-\frac{1}{\sqrt{g_1 g_2}} < \rho < \frac{1}{\sqrt{g_1 g_2}} \quad . \tag{20}$$

The message for geologists is that the allowable range of background correlation is restricted (sometimes severely) by assignment of within cluster correlations. A coherent assessment scheme must account for this constraint. Consider Table 1, Table 1 green and blue matrices are examples of intra-class correlation matrices for which computation of upper and lower bounds on background correlation ρ are particularly simple. The proof relies on elementary properties of an intra-class correlation matrix:

(a) An $(N \times N)$ intra-class correlation matrix with correlation coefficient θ is *PDS* *iff* $-\frac{1}{N-1} < \theta < 1$.

(b) The sum of elements of the inverse of an $(N \times N)$ intra-class correlation matrix with correlation coefficient θ is $N/(1+(N-1)\theta)$.

The sum of elements of the inverse of the green matrix in Table 1 is $g_1 = 1.346$ and sum of elements of the inverse of the blue matrix is $g_2 = 1.471$. Pairwise background correlation ρ between clusters is restricted to lie in $(-0.711, 0.711)$. If the green matrix is replaced with a

(7×7) identity matrix and the blue matrix is replaced with a (5×5) identity matrix then ρ is restricted to lie in $(-0.169, 0.169)$.

5.2 Background Correlation

Partition basic unit magnitudes into K clusters and write the vector of uncertain basic unit magnitudes as $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_K)^t$. Redefine the variance matrix of $\mathbf{X}$ to be

$$Var(\mathbf{X}) = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \cdots & \mathbf{V}_{1K} \\ \mathbf{V}_{21} & \mathbf{V}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{V}_{K-1,K} \\ \mathbf{V}_{K1} & \cdots & \mathbf{V}_{K,K-1} & \mathbf{V}_{KK} \end{bmatrix}, \mathbf{V}_{ii} = Var(\mathbf{X}_i) \text{ and } \mathbf{V}_{ij} = Cov(\mathbf{X}_i, \mathbf{X}_j). \qquad (21)$$

Interpret $\mathbf{V}_{ii}$ as the variance matrix assigned to magnitudes of geologically similar basic units assigned to the i^{th} cluster. In (21) $\mathbf{V}_{ij}$ is the covariance of pairs of elements, one in cluster i and the other in cluster $j, i \neq j$.

If K is small to moderate and no single cluster is large, geologists can often provide coherent assessment of elements of each $\mathbf{V}_{ii}$. However, even when the number of clusters is small to moderate, the number of pairwise co-variances can be large enough to make direct subjective assessment of dependencies between units in distinct clusters impractical. Assignment of a common background correlation to all unit magnitudes as in Table 1 or to unit magnitudes as in the Circum-Arctic study is one tactic for dodging this difficulty. Define $\mathbf{D}_i^{1/2}$ to be a diagonal matrix with diagonal elements composed of positive square roots (standard deviations) of diagonal elements of $\mathbf{V}_{ii}$ and

$$\mathbf{D}^{1/2} = \begin{bmatrix} \mathbf{D}_1^{1/2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0}^t & \mathbf{D}_2^{1/2} & \vdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0}^t & \cdots & \mathbf{0}^t & \mathbf{D}_K^{1/2} \end{bmatrix}. \tag{22}$$

The correlation matrix of $\mathbf{X}$ is

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1K} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{C}_{K-1,K} \\ \mathbf{C}_{K1} & \cdots & \mathbf{C}_{K,K-1} & \mathbf{C}_{KK} \end{bmatrix} = \mathbf{D}^{-1/2}\mathbf{V}\mathbf{D}^{-1/2}. \tag{23}$$

Within cluster correlation matrices are $\mathbf{C}_{kk} = \mathbf{D}_k^{-1/2}\mathbf{V}_{kk}\mathbf{D}_k^{-1/2}, k = 1,...,K$ and $\mathbf{C}_{ij} = \mathbf{D}_i^{-1/2}\mathbf{V}_{ij}\mathbf{D}_j^{-1/2}$, $i, j = 1,...,K, j \neq i$ are between cluster correlation matrices. Assigning special structure to $\mathbf{C}$ reduces the assessment burden in return for restricting allowable ranges of some pairwise correlations. For example, assume that all elements of each $\mathbf{C}_{ij}, i \neq j$ equal a common background correlation ρ. Define $\mathbf{1}_i^t$ to be a $(n(i) \times 1)$ vector of 1s and set

$$\mathbf{C}_{ij} = \mathbf{1}_i\mathbf{1}_j^t \times \rho \; i, j = 1,...N, i \neq j \; . \tag{24}$$

The correlation matrix, Eq. (23) is then

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{1}_1\mathbf{1}_2^t\rho & \cdots & \mathbf{1}_1\mathbf{1}_K^t\rho \\ \mathbf{1}_2\mathbf{1}_1^t\rho & \mathbf{C}_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{1}_{K-1}\mathbf{1}_K^t\rho \\ \mathbf{1}_K\mathbf{1}_1^t\rho & \cdots & \mathbf{1}_{K-1}\mathbf{1}_K^t\rho & \mathbf{C}_{KK} \end{bmatrix}. \tag{25}$$

The Circum-Arctic study employs a version of (25) with a small number of off block diagonal correlations assigned values different from ρ. In order for $\mathbf{C}$ as in (25) to be PDS the correlation coefficient ρ must lie in an interval (ρ^-, ρ^+) with ρ^- the largest negative root and ρ^+ the

smallest positive root of a polynomial $P(\rho)$ of degree K whose coefficients are composed of elementary symmetric functions of $g_1,...,g_K$ with g_k the sum of elements of $\mathbf{C}_{kk}^{-1}$. Alternatively, on applying a similarity transform to $\mathbf{C}$ that maps it into block diagonal form, ρ^- is the largest negative eigenvalue and ρ^+ is the smallest positive eigenvalue of a $(K \times K)$ matrix appearing on the diagonal of transformed $\mathbf{C}$ (Kaufman (2016)).

5.3 Aggregation of Clusters

To illustrate how aggregation affects order relations between background correlation and correlations of aggregates partition a $(N \times 1)$ vector $\mathbf{X}$ of basic unit magnitudes as $\mathbf{X} = (\mathbf{X}_1^t, \mathbf{X}_2^t)^t$ and let

$$Var(\mathbf{X}) = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \tag{26}$$

as in Eq. (17). Define $\mathbf{D}_1 = diag(v_{11},...,v_{nn})$, $\mathbf{D}_2 = diag(v_{n+1,n+1},...,v_{NN})$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0}^t & \mathbf{D}_2 \end{bmatrix} \tag{27}$$

and factor $Var(\mathbf{X})$ as

$$Var(\mathbf{X}) = \mathbf{D}^{1/2}\mathbf{C}\mathbf{D}^{1/2} \text{ with conformable } \mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}. \tag{28}$$

Assume that pairwise correlations between elements of $\mathbf{X}_1$ and elements of $\mathbf{X}_2$ are $\mathbf{C}_{12} = \mathbf{1}_1\mathbf{1}_2^t\rho$ as in Eq. (25).

Assertion 5: Provided that $X_1,...,X_N$ are positively correlated and elements of $\{X_1,...,X_n\}$ and $\{X_{n+1},...,X_N\}$ possess common (background) correlation ρ, the pairwise correlation of $S_1 = X_1 + ... + X_n$ and $S_2 = X_{n+1} + ... + X_N$ is larger than ρ.

(a) The pairwise correlation of S_1 and S_2 is

$$\frac{\rho}{\sqrt{\bar{c}_n + (1-\bar{c}_n)\bar{\bar{f}}_n} \times \sqrt{\bar{\bar{c}}_{N-n} + (1-\bar{\bar{c}}_{N-n})\bar{\bar{\bar{f}}}_{N-n}}} \quad . \tag{29}$$

Here, parameters $\bar{c}_n$ and $\bar{\bar{c}}_{N-n}$ are weighted correlations defined in Eq. (35) and Eq. (37),

$$\bar{\bar{f}}_n = \frac{v_{11} + ... + v_{nn}}{(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2} \text{ and } \bar{\bar{\bar{f}}}_{N-n} = \frac{v_{n+1,n+1} + ... + v_{NN}}{(\sqrt{v_{n+1,n+1}} + ... + \sqrt{v_{NN}})^2}. \tag{30}$$

(b) If variances v_{ii} are bounded away from zero and are finite, background correlation ρ is less than the geometric mean of $\bar{c}_n$ and $\bar{\bar{c}}_{N-n}$:

$$\rho < (\bar{c}_n \times \bar{\bar{c}}_{N-n})^{1/2}(1 + O(\frac{1}{N})). \tag{31}$$

If pairwise correlations within clusters are bounded away from zero then $\bar{c}_n$ and $\bar{\bar{c}}_{N-n}$ are both $O(1)$. If Eq. (20) implies that within cluster correlations are $O(\frac{1}{N})$ then $\bar{c}_n$ and $\bar{\bar{c}}_{N-n}$ are $O(\frac{1}{N})$.

Proof: Use the identity

$$(\sqrt{v_1} + ... + \sqrt{v_n})^2 = v_1 + ... + v_n + \sum_{i \neq j} \sqrt{v_i v_j}. \tag{32}$$

Apply it to the variance of a sum of *rvs* with variance matrix $(n \times n)\,\mathbf{V}_{11} > \mathbf{0}$ possessing diagonal elements $v_{11},...,v_{nn}$. Equation (32) implies the inequality

$$v_{11} + ... + v_{nn} + \sum_{i \neq j} c_{ij} \sqrt{v_{ii} v_{jj}} < (\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2 . \quad . \tag{33}$$

Let $\sum^{*} \equiv \sum_{i,j=1,...,n,\, i \neq j}$ and write

$$Var(S_1) = v_{11} + v_{nn} + \sum^{*} c_{ij} \sqrt{v_{ii} v_{jj}} . \tag{34}$$

Use $\sum^{*} \sqrt{v_{ii} v_{jj}} = (\sqrt{v_{11}} + ... \sqrt{v_{nn}})^2 - (v_{11} + ... + v_{nn})$ to construct a weighted correlation

$$\bar{c}_n = \sum^{*} c_{ij} \frac{\sqrt{v_{ii} v_{jj}}}{(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2 - v_{11} - ... - v_{nn}} . \tag{35}$$

Then $\sum^{*} c_{ij} \sqrt{v_{ii} v_{jj}} = \bar{c}_n \times [(\sqrt{v_{11}} + ... \sqrt{v_{nn}})^2 - (v_{11} + ... + v_{nn})]$ so the variance of S_1 is representable as

$$Var(S_1) = (1 - \bar{c}_n) \times (v_{11} + ... + v_{nn}) + \bar{c}_n \times (\sqrt{v_{11}} + .. \sqrt{v_{nn}})^2 . \tag{36}$$

Treat $Var(S_2)$ similarly. Define $\sum_{k,l=n+1,...,N,\, k \neq l} = \Sigma^{**}$ and

$$\bar{\bar{c}}_{N-n} = \sum^{**} c_{kl} \frac{\sqrt{v_{kk} v_{ll}}}{(\sqrt{v_{n+1,n+1}} + ... + \sqrt{v_{NN}})^2 - v_{n+1,n+1} - ... - v_{NN}} \tag{37}$$

so that

$$Var(S_2) = (1 - \bar{\bar{c}}_{N-n}) \times (v_{n+1,n+1} + ... + v_{NN}) + \bar{\bar{c}}_{N-n} \times (\sqrt{v_{n+1,n+1}} + ... \sqrt{v_{NN}})^2 . \tag{38}$$

The pairwise correlation of S_1 and S_2 is

$$Corr(S_1, S_2) = \frac{(\sqrt{v_{11}} + ... + \sqrt{v_{nn}}) \times (\sqrt{v_{n+1,n+1}} + ... + \sqrt{v_{NN}})}{\sqrt{Var(S_1)Var(S_2)}} \times \rho . \tag{39}$$

Using Eq. (36)

$$\frac{\sqrt{v_{11}} + ... + \sqrt{v_{nn}}}{\sqrt{Var(S_1)}} = [(\bar{c}_n + (1 - \bar{c}_n) \times \frac{v_{11} + ... + v_{nn}}{(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2}]^{-1/2} . \tag{40}$$

As $\bar{c}_n \in (0,1)$ and $(v_{11} + ... + v_{nn}) / (\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2 < 1$ the denominator on the RHS of Eq. (40) is less than one. Treat $Var(S_2)$ similarly. Taken together equations (36), (38), (39) and Eq. (40) yield

$$\rho < [\bar{c}_n + (1 - \bar{c}_n)\bar{\bar{f}}_n]^{1/2} \times [\bar{\bar{c}}_{N-n} + (1 - \bar{\bar{c}}_{N-n})\bar{\bar{\bar{f}}}_{N-n}]^{1/2} . \tag{41}$$

Suppose that variances are bounded away from zero and are restricted to be finite so that there exists small $\varepsilon > 0$ and large $B > 0$ such that $\varepsilon < v_{ii} < B$. Then

$$\frac{\varepsilon}{nB} \le \frac{v_{11} + ... + v_{nn}}{(\sqrt{v_{11}} + ... + \sqrt{v_{nn}})^2} \le \frac{B}{n\varepsilon} \tag{42}$$

so $\bar{\bar{f}}_n$ is $O(\frac{1}{n})$ and $\bar{\bar{\bar{f}}}_{N-n}$ is $O(\frac{1}{N-n})$. In the uniform asymptotic regime $N \to \infty, \frac{n}{N} \to \alpha$ with α bounded away from zero and one $\bar{\bar{f}}_{N\alpha} \to 0$ and $\bar{\bar{\bar{f}}}_{N(1-\alpha)} \to 0$ leading to Eq. (31).

A simple example is informative. Suppose that $Var(\mathbf{X}_1) = Corr(\mathbf{X}_1)$ is the (7×7) intra-class correlation green matrix in Table 1 with $\theta_1 \equiv 0.7$, $Var(\mathbf{X}_2) = Corr(\mathbf{X}_2)$ is the (5×5) blue intra-

class correlation matrix with $\theta_2 \equiv 0.6$ and $Cov(\mathbf{X}_1, \mathbf{X}_2) = Corr(\mathbf{X}_1, \mathbf{X}_2)$ is the black matrix with common elements $\rho = 0.4$. Direct computation yields

$$Corr(S_1, S_2) = \frac{\rho}{\sqrt{\theta_1 + \frac{1-\theta_1}{n}} \times \sqrt{\theta_2 + \frac{1-\theta_2}{N-n}}}. \tag{43}$$

The allowable range of θ_1 is $(-\frac{1}{n-1}, 1)$ and that of θ_2 is $(-\frac{1}{N-n-1}, 1)$. For θ_1 and θ_2 in their allowable ranges $\theta_1 + \frac{1-\theta_1}{n}$ and $\theta_2 + \frac{1-\theta_2}{N-n}$ are positive. Because the sum of elements of the inverse of the green matrix in Table 1 is $g_1 = n/(1+(n-1)\theta_1)$ and the sum of elements of the blue matrix is $g_2 = (N-n)/(1+(N-n-1)\theta_2$, Assertion 4 says that the correlation coefficient $\rho(>0)$ must be less than the denominator in (43) in order for $Var(\mathbf{X})$ with $\mathbf{X} = (\mathbf{X}_1^t, \mathbf{X}_2^t)$ to be positive definite. (The Cauchy-Schwartz inequality says the same). In this example $n = 7, N-n = 5, \bar{c}_n = \theta_1 = 0.7, \bar{\bar{c}}_{N-n} = \theta_2 = 0.6$, $\overline{f_n} = 1/n = 1/7$ and $\overline{\overline{f}}_{N-n} = 1/(N-n) = 1/5$. For $\rho = 0.4$ $Corr(S_1, S_2) = 0.563$. The allowable range of ρ is $(-0.711, 0.711)$.

Consider an alternative partition of the twelve basic units in Table 1 into two subsets with labels $\{1,2,3,4,5,6,7,8,9\}$ and $\{10,11,12\}$. This partition "splits" clusters in such a way that common background correlations of 0.4 in Table 1 appear in the correlation matrix for units labelled $\{1,2,3,4,5,6,7,8,9\}$ along with correlations 0.7. The set of all pairwise correlations between elements of $\{10,11,12\}$ and elements of $\{1,2,3,4,5,6,7,8,9\}$ are no longer identical so $S_1^* = X_1 + X_2 + ... + X_9$ and $S_2^* = X_{10} + X_{11} + X_{12}$ are not block diagonal aggregates. In this particular case the pairwise correlation of S_1^* and S_2^* is 0.635, substantially greater than

background correlation 0.4. In general, when a partition of basic unit magnitudes splits clusters possessing common background correlations as in the above example the pairwise correlation between resulting sums can be greater than, equal to or less than background correlation.

6 Assessment tradeoffs

6.1 Parsimony

Direct assessment of second moments of *N* unit magnitudes requires appraisal of N variances and $\frac{1}{2}N(N-1)$ pairwise correlations. Partitioning aggregate units into a small number of clusters and directly assessing correlations between sums of basic unit magnitudes in each cluster (multiple stage aggregation) is an attractive alternative because it modulates the tyranny of large numbers. An example is direct assessment of correlations between pairs of sums of oil equivalent in each of several petroleum plays instead of between individual prospects and accumulations. To a geologist assigned the task of assessing co-variability among basic unit magnitudes this sounds like a magically simple recipe! The number of pairwise correlation coefficients decreases at the expense of requiring subjective appraisal of co-variability of sums of oil equivalents. Aggregation of $X_1,...,X_N$ to $S_1,...,S_K$, $K<N$ requires specification of $K(K-1)/2$ pairwise correlation coefficients and K variances. The reduction in number of parameters to assess is beguiling! For example, a (12×12) variance matrix of basic unit magnitudes $X_1,...,X_{12}$ requires specification 12 variances and 66 pairwise co-variances. If $X_1,...,X_{12}$ are partitioned into two subsets $\{X_1,...,X_7\}$ and $\{X_8,...,X_{12}\}$, the variance matrix for sums $S_1=X_1+...,X_7$ and $S_2=X_8+...,X_{12}$ possesses only three parameters. Asking a geologist to assess $Var(S_1), Var(S_2)$ and $Cov(S_1,S_2)$ in place of parameters of the variance matrix

of $X_1,...,X_{12}$ greatly reduces the assessment burden but shifts focus away from properties of basic unit magnitudes. A modeling tactic that trims the number of parameters to assess is to assign common pairwise correlation of oil equivalent magnitudes to members of each petroleum play in a sample frame and to specify a common background correlation between oil equivalent magnitudes in distinct plays—as in Table 1 for example.

Because variances and co-variances of aggregates of $X_1,...,X_N$ are functionally dependent on all variances and co-variances of $X_1,...,X_N$ introduction of multiple levels of aggregation reduces the number of parameters to assess but increases the number of constraints on second moments of aggregates. Assessment schemes must take these features of aggregates into account.

6.2 Elicitation of Dependencies and Correlation

When measureable data available to estimate oil and gas depositional model parameters are not available the only way to proceed is to elicit geologists' judgments about parameters and dependencies (Meyer and Booker 2001; O'Hagan et al. 2006; Delfiner and Barrier 2008; Daneshkhah and Oakley 2010). Geologic analogy (a qualitative measure of similarity) plays an important role here. Choice of which analogy is highly subjective, adding a layer of complexity to the assessment process. Team effects occur often: correlations among basic units in distinct areas assessed by a particular team are often larger than correlations between units assessed by that team and units assessed by a different team. In addition to these assessment issues, subjective appraisal of pairwise co-variability by elicitation of judgments about pairwise correlations deserves particular attention. Appraisal of the impact of subjective assessment biases is important. If basic units are probabilistically dependent, assessment error at one unit can propagate to assessment errors at other units.

The Pearson correlation coefficient $-1 < \rho < 1$ is a measure of the strength of linear association between two uncertain quantities. Although it can be computed for any pair of uncertain quantities whose joint distribution is known, an estimate of it computed from observed data is neither robust nor resistant to outliers (Wilcox 2016) and if not carefully interpreted can be misleading (Anscombe 1973). Here pairwise correlations are not estimated from data so estimation robustness is not an issue. However, use of pairwise correlation as a measure of dependence of one random variable Y on another random variable X can be misleading in other ways. How to interpret the meaning of ρ depends on the particular joint probability law governing X and Y. If X and Y are bivariate normal the expectation $E(Y|X)$ of Y given X is a linear function of X so ρ is a sensible measure of the elasticity (variation of) Y with respect to X as well as of the dispersion of Y around the regression line $E(Y|X) = a + bX$. If X and Y are bivariate lognormal $E(Y|X)$ is no longer a linear function of the pairwise correlation of X and Y. More generally, the pairwise correlation between functions of two bivariate Normal *rvs* is not a robust measure of dependency.

It is more natural for geologists to think about how the magnitude X_i of basic unit i varies with variations in X_j rather than how $\ln X_i$ varies as $\ln X_j$ varies. For $i \neq j$ suppose that the $(i, j)^{th}$ element c_{ij} of $\mathbf{C}$ is the pairwise correlation between basic unit magnitudes. In general $Corr(X_i, X_j)$ is not equal to $Corr(\ln X_i, \ln X_j)$. However, when $\mathbf{X}_N$ is multivariate lognormal with $Var(\ln X_i) = \sigma_{ii}^2, i = 1, ..., N$ and $Corr(\ln X_i, \ln X_j) = r$ fixed, for small $\sigma_{ii}^2, i = 1, ..., N$ $Corr(X_i, X_j) \approx r$. A protocol designed to elicit geologists' judgments about degrees of dependencies among basic unit magnitudes assumed to be lognormal must take into account

these facts. A review of some analytical methods for modeling dependencies that go beyond pairwise correlation and its cousins appears in Kaufman (2018).

6.3 Principal Conclusions

Any method of aggregation of basic unit magnitudes obeying the rules of probability leads to the same distribution of the sum of all unit magnitudes.

If not carefully policed, personal (subjective) judgments by geologists elicited at distinct levels of aggregation may or may not be coherent and may or may not lead to a distribution for the sum of all basic unit magnitudes identical to that computed by direct summation of all of them. This is an implementation not a mathematical problem. Multiple stage aggregation requires fewer judgmental assessments about fewer parameters. The tradeoff is that multiple stage aggregation directs geologists' subjective probability assessments away from primitive geological attributes underpinning properties of basic unit magnitudes.

Probabilistic aggregation of resources that incorporates expert judgment is coherent if and only if judgments adhere to the rules of probability. Resolution of many issues that plague assessment practice remain to be studied and resolved.

Acknowledgments The U. S. Geological Survey requires a preliminary internal review before any paper can be published in a scientific journal (http://pubs.usgs.gov/circ/1367/). We wish to thank Emil Attanasi, David Root and Peter Warwick for their insightful suggestions.

References

Anscombe FJ (1973) Graphics in statistical analysis. The American Statistician 27(1): 17–21

Blondes MS, Brennan ST, Merrill MD, Buursink ML, Warwick PD, Cahan SM, Cook TA, Corum MD, Craddock WH, DeVera CA, Drake RM, Drew LJ, Freeman PA, Lohr CD, Olea RA, Roberts-Ashby TL, Slucher .R, Varela BA (2013a) National assessment of geologic carbon dioxide storage resources—Methodology implementation: U.S. Geological Survey Open-File Report 2013-1055, 26 p. http://pubs.usgs.gov/of/2013/1055/OF13-1055.pdf

Blondes MS, Schuenemeyer JH, Drew LJ, Warwick PD (2013b) Probabilistic aggregation of individual assessment units in the U.S. Geological Survey national CO_2 sequestration assessment: Energy Procedia 37:5110−5117.

Blondes MS, Schuenemeyer JH, Olea RA, Drew LJ (2013c) Aggregation of carbon dioxide sequestration storage assessment units. Stochastic Environmental Research and Risk Assessment 27:1839−1859

Carter PJ, Morales E (1998) Probabilistic addition of gas reserves within a major gas project. Paper presented at the Society of Petroleum Engineers Asia Pacific Oil and Gas Conference and Exhibition, 8 p. SPE paper 50113

Collett T (2008) Assessment of Gas Hydrates on the North Slope, Alaska, 2008. US Geological Survey Fact 2008-3073, 4 p. https://pubs.usgs.gov/fs/2008/3073/pdf/FS08-3073_508.pdf

Crovelli RA, Balay RH (1991) A microcomputer program for energy assessment and aggregation using the triangular probability distribution. Computers & Geosciences 17(2):197–225

Daneshkhah A, Oakley JE (2010) Eliciting multivariate probability distributions. In: Rethinking risk measurement and reporting: Volume I, Böcker K (ed). Risk Books, London

Delfiner P, Barrier R (2008) Partial probabilistic addition: A practical approach for aggregating resources. SPE Reservoir Evaluation and Engineering 11(2):379–386. SPE paper 90129

Horn RA, Johnson CR (2013) Matrix analysis. Second edition, Cambridge University Press, New York, 643 p

Kaufman GM (2016) Generalizations of intra-class correlation matrices (unpublished working paper)

Kaufman GM (2018 forthcoming) Properties of Sums of Geologic Random Variables Ch. 5 in Handbook of Mathematical Geosciences: Fifty Years of IAMG. Ed. B.S. Daya Sagar, Quiming Cheng, Fritz Agterberg

Klett TR, Gautier DL (2009) Assessment of undiscovered petroleum resources of the Barents Sea. U.S. Geological Survey Fact Sheet 2009-3037, 4 p. http://pubs.usgs.gov/fs/2009/3037/pdf/FS09-3037.pdf

Meyer MA, Booker JM (2001) Eliciting and analyzing expert judgement: A practical guide. ASA-SIAM Series on Statistics and Applies Probabilities, Alexandria, VA, 459 p

Miller, BM, Thomsen, HL, Dolton, GL, Coury, AB, Hendricks, TA, Lennartz, FE, Powers, R., Sable, EG, Varnes, KI (1975) Geological estimates of undiscovered oil and gas resources in the United States. United States Geological Survey Circular 725, 78 pages, 3 maps

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: Eliciting experts' probabilities. John Wiley & Sons, Chichester, UK, 321 p

Pike R (2008) How much oil is really there? Making correct statistics bring reality to global planning. Significance 5:149–152

Schuenemeyer JH (2005) Methodology for the 2005 USGS assessment of undiscovered oil and gas resources, Central North Slope, Alaska. U.S. Geological Survey Open-File Report 2005-1410, 82 p. https://pubs.usgs.gov/of/2005/1410/of2005-1410.pdf

Schuenemeyer JH, Gautier DL (2010) Aggregation methodology for the Circum-Artic resource appraisal. Mathematical Geosciences 42(5):583–594

U.S. Geological Survey Geologic Carbon Dioxide Storage Resources Assessment Team (2013) National assessment of geologic carbon dioxide storage resources-Results: U.S. Geological Survey Circular 1386, 41 p. http://pubs.usgs.gov/circ/1386/pdf/circular1386.pdf

Van Elk JF, Gupta R (2010) Probabilistic aggregation of oil and gas field resource estimates and project portfolio analysis. SPE Reservoir Evaluation & Engineering 13(1):72–81. SPE paper 116395

Wilcox, RR (2016) Introduction to robust estimation and hypothesis testing. Academic Press, fourth edition, 786 p

Table 1 Basic unit magnitude correlation matrix

1											
0.7	1										
0.7	0.7	1									
0.7	0.7	0.7	1								
0.7	0.7	0.7	0.7	1							
0.7	0.7	0.7	0.7	0.7	1						
0.7	0.7	0.7	0.7	0.7	0.7	1					
0.4	0.4	0.4	0.4	0.4	0.4	0.4	1				
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.6	1			
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.6	0.6	1		
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.6	0.6	0.6	1	
0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.6	0.6	0.6	0.6	1

Table 2 Basic unit variance matrix

	1	2	3	4	5	6	7	8	9	10	
1	670	606	595	633	508	460	604	433	346	329	3
2		1117	769	817	656	594	780	559	447	425	4
3			1080	803	645	584	767	550	440	418	4
4				1219	685	621	815	584	467	444	4
5					786	498	655	469	375	356	3
6						645	593	425	340	323	3
7							1113	558	446	424	4
8								1751	839	798	8
9									1118	637	6
10										1009	6
11											1

12

Table 3 Variance and correlation matrices for clusters {1,2,3,4}, {5,6,7}, {8,9,10}, {11,12}

{1,2,3,4}	11948	7188	4809	3389	1	0.864	0.507	0.474
{5,6,7}	7188	5787	3252	2352	0.864	1	0.493	0.473
{8,9,10}	4809	3252	7523	4376	0.507	0.493	1	0.771
{11,12}	3389	2352	4376	4278	0.474	0.473	0.771	1

Table 4 Variance and correlation matrices for clusters {1,2,3,4,5,6,7} and {8,9,10,11,12}

{1,2,3,4,5,6,7}	32112	13802	1	0.537
{8,9,10,11,12}	13802	20554	0.537	1

Table 5 Properties of marginal lognormal distributions. St. Dev = standard deviation

	Unit Number											
	1	2	3	4	5	6	7	8	9	10	11	12
Mean	27.6	42.3	61.8	73.8	32.5	35.2	49.0	55.9	67.4	40.2	45.5	59.8
St Dev	25.9	33.4	32.9	34.9	28.0	25.4	33.4	41.8	33.4	31.8	33.4	40.8
Median	20.16	33.18	54.61	66.70	24.60	28.54	40.49	44.77	60.35	31.56	36.65	49.45
Mode	10.73	20.43	42.59	54.51	14.11	18.77	27.65	28.70	48.44	19.43	23.80	33.77
0.9 Fractile	56	81	103	119	64	65	89	105	110	77	85	109
0.1 Fractile	9.0	17.8	39.7	51.5	12.1	16.7	24.8	25.3	45.5	17.0	21.1	30.3
μ	3.004	3.502	4.000	4.200	3.203	3.351	3.701	3.801	4.100	3.452	3.601	3.901

σ	0.794	0.696	0.499	0.449	0.745	0.647	0.618	0.667	0.469	0.696	0.657	0.618
σ^2	0.631	0.485	0.249	0.202	0.556	0.419	0.381	0.445	0.220	0.485	0.432	0.381

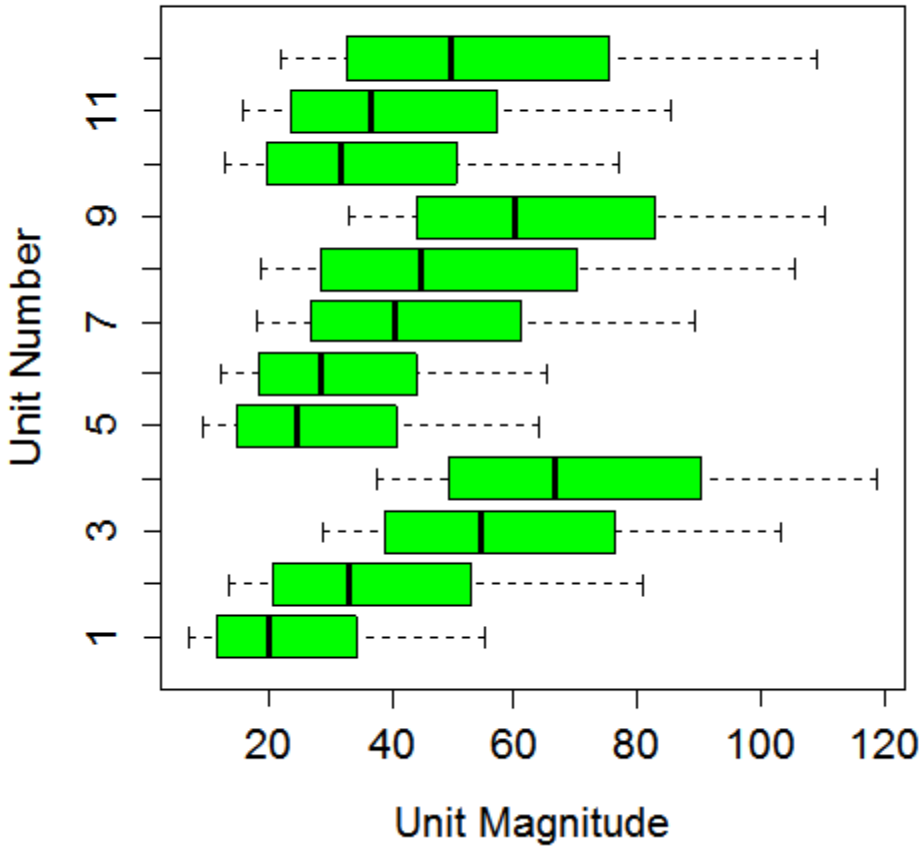

Fig. 1 Basic unit magnitude boxplots

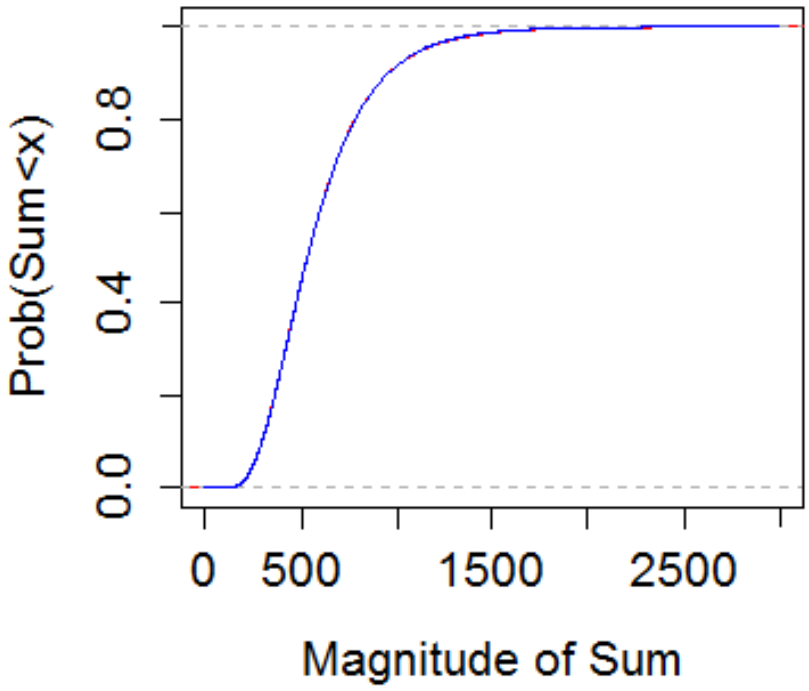

Fig. 2 Overlay of empirical and lognormal fit to cumulative distribution function of the sum of all basic unit magnitudes

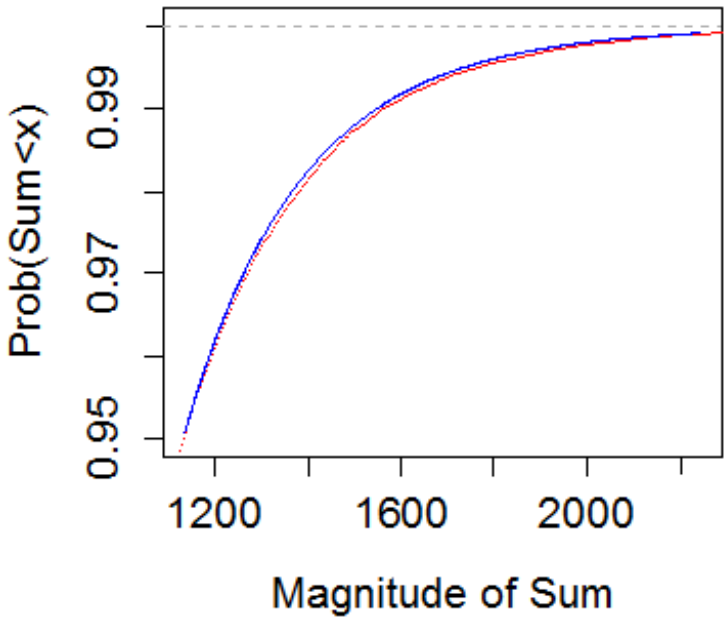

Fig. 3 Overlay from 0.95^{th} to 0.999^{th} fractiles

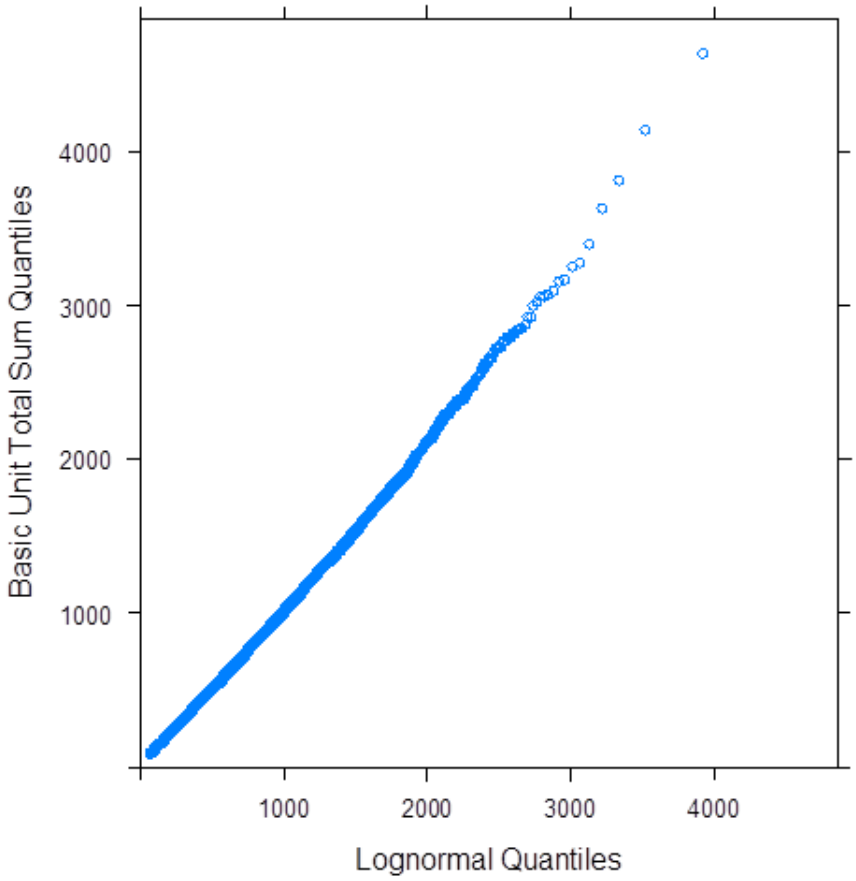

Fig. 4 Lognormal Q-Q plot of sum of all basic unit magnitudes

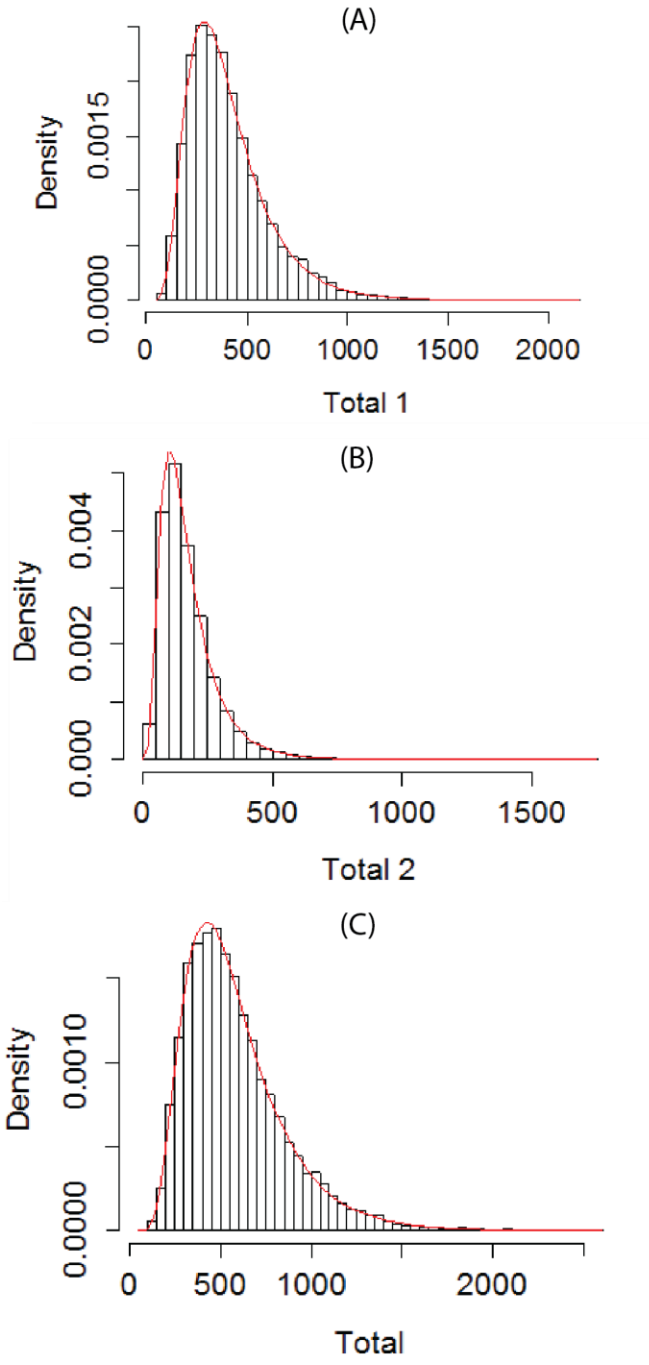

Fig. 5 Histograms and fitted lognormal distributions: (A) aggregation of basic units 1−7 (B) aggregation of basic units 8−12 (C) sum of (A) and (B)

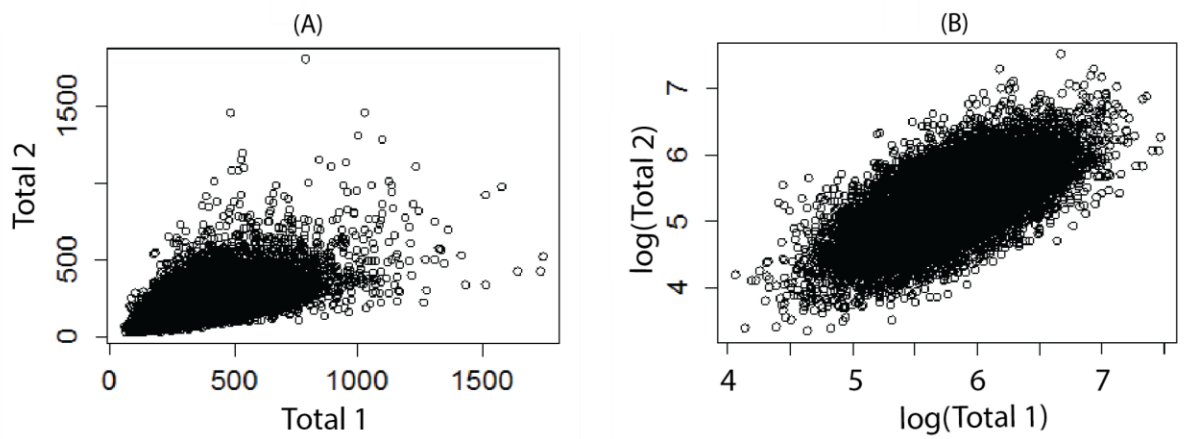

Fig. 6 Scatterplot of partial aggregations of basic units 1−7 and 8−12: (A) original space; (B) logarithmic space

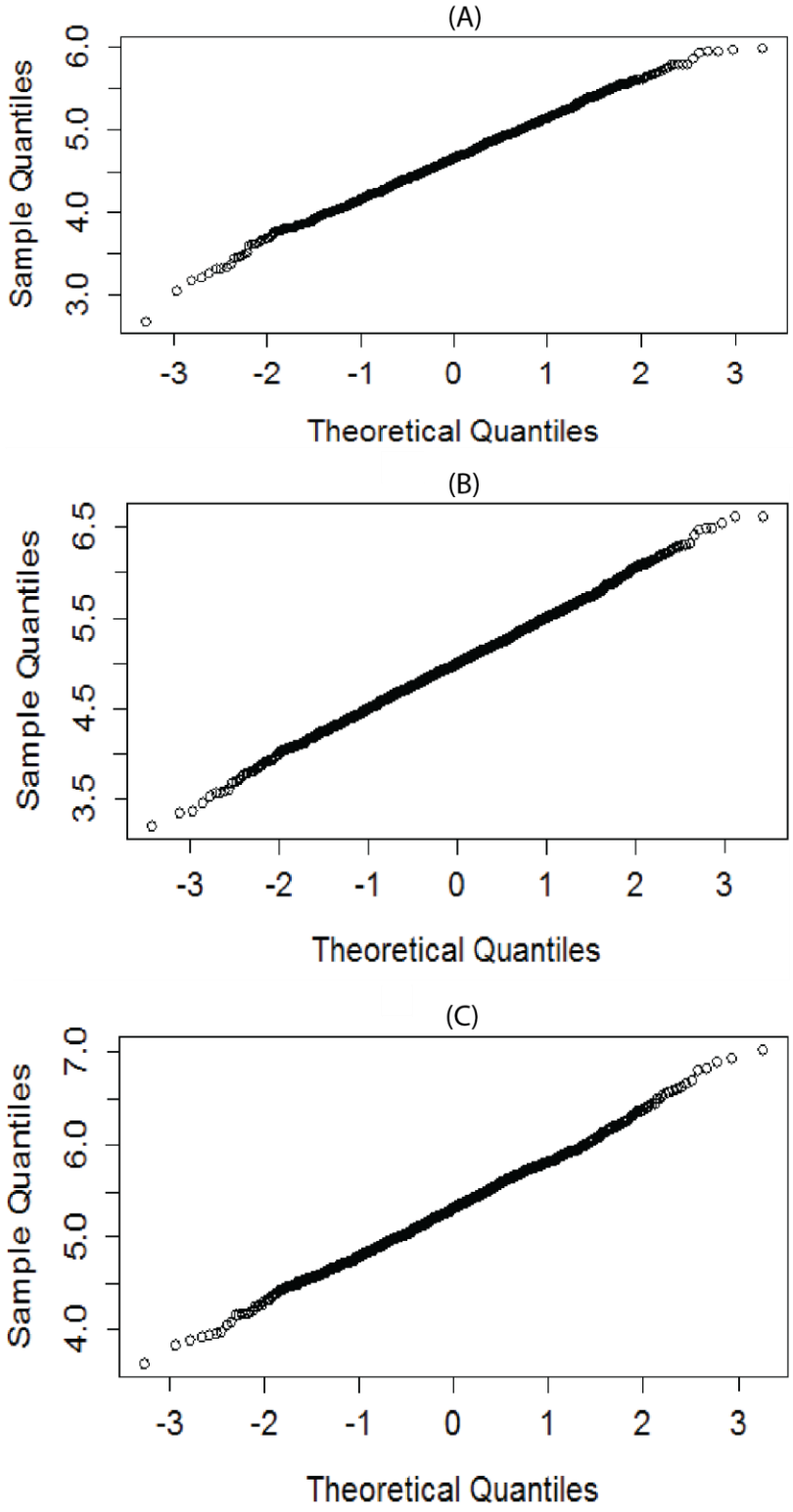

Fig. 7 Q-Q slices of ln(Total2): (A) near −1 standard deviation; (B) near center; (C) near +1 standard deviation.