How Accurate Are Expert Estimations of Correlation?

Michael Beer
Institute for Risk and Reliability
Callinstraße 34, Leibniz Universität Hannover
30167 Hannover, Germany, and
Institute for Risk and Uncertainty
University of Liverpool
Liverpool L69 3BX, United Kingdom, and
International Joint Research Center
for Engineering Reliability
and Stochastic Mechanics (ERSM)
Tongji University
1239 Siping Road, Shanghai 200092,P.R.China
beer@irz.uni-hannover.de

Zitong Gong
Francisco Alejandro Diaz De La O
Institute for Risk and Uncertainty
School of Engineering
University of Liverpool
Liverpool L69 3BX, United Kingdom
Zitong.Gong@liverpool.ac.uk
f.a.diazdelao@liverpool.ac.uk

Vladik Kreinovich
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu

***Abstract*—In many practical situations, it is important to know the correlation between different quantities – finding correlations helps to gain insights into various relationships and phenomena, and helps to inform analysts. Often, there is not enough empirical data to experimentally determine all possible correlations. In such cases, a natural idea is to supplement this situation with expert estimates. Expert estimates are rather crude. So, to decide whether to act based on these estimates, it is desirable to know how accurate are expert estimates. In this paper, we propose several techniques for gauging this accuracy.**

***Index Terms*—correlation, expert estimates, uncertainty, fuzzy sets, vagueness**

I. Formulation of the Problem

Correlations are important. How can we cure diseases? How can we prevent diseases? Often, we do not know what causes a disease, and we do not know what helps against this disease. In such situations, we collect the data about the patients and try to find the solutions by analyzing the data.

If a certain disease is strongly correlated with, e.g., smoking, then probably either smoking causes this disease or, alternatively, smoking weakens the body's natural defenses against this disease. If a disease is correlated with the presence of a certain variant gene (or, more generally, with a combination of variant genes), then people with this combination should regularly check again this disease and/or perform some additional preventive measures against this disease.

Similarly, in geosciences, if oil locations are strongly correlated with certain types of geological structures, then we should actively look for oil in an area where such structures are present. In many application areas, correlations are important.

It is often not possible to find correlations experimentally, so we need to rely on experts. In the ideal world, we should be able to determine all the correlations from experiments. In practice, however, it is not always possible.

An ideal way to estimate the correlation between the two quantities x and y is to have a sample of data points in which only x changes – all other parameters remain the same – and we are interested in how y changes depending on the change in x. This may be possible if we analyze how a given photosensor reacts to temperature, but it is not realistic to expect many situations in which patients are identical in almost all characteristics – expect that some of them do not smoke, some smoke lightly, and some smoke heavily. Similarly, in geosciences, we cannot expect situations in which all the geological characteristics are identical except for one feature.

In most practical situations, we have many factors affecting each situation, and there is often not enough data points to separate the effect of these factors. In situations when we cannot determine all the correlations empirically, a natural idea is to ask experts. Experts can often provide their estimates of correlation between different quantities.

What type of information can experts provide? First, experts can provide us with numerical estimates for the correlation between the given n quantities $x_1, \ldots, x_n$. In other words, for every $i \neq j$, the experts will provide us with an estimate a_{ij} of the correlation between the quantities x_i and x_j. Due to the properties of correlation, we always have

$$a_{ij} = a_{ji}.$$

Since correlation between a quantity and itself is 1, all diagonal elements of the correlation matrix are 1s. As a result, by adding 1s on the diagonal, we get a symmetric matrix a_{ij} in which all diagonal elements are equal to 1.

While experts can (and do) produce numbers, they are not 100% confident in these numbers. As a next step, it is therefore

reasonable to capture this expert uncertainty. The experts often describe their uncertainty in terms of imprecise (fuzzy) words from natural language, such as "close to", "much smaller than", "about". To describe the expert's uncertainty, it is therefore reasonable to use techniques specifically developed for translating such statements into numerical form – namely, the techniques of fuzzy logic; see, e.g., [6], [9], [11]. As a result, each estimate a_{ij} is no longer a number – it is a *fuzzy number*, describing the expert's uncertainty.

How accurate are expert's estimates of correlation? As we have mentioned, expert's estimates are approximate. How accurate are they? It is important to know this since it will determine whether we should act on these correlations or maybe perform more experiments.

For example, if an expert claims that there is a 70% correlation between a certain gene and a disease, and the expert's accuracy is $\pm 20\%$, this means that we are confident that there is a strong positive correlation. On the other hand, if an expert estimates the correlation at 20% and his/her accuracy is $\pm 30\%$, then maybe there is no positive correlation at all – or even a small negative correlation.

It is therefore important to determine how accurate are the expert's estimates of correlation. This paper proposes a method to do this.

II. The Main Idea and Resulting Precise Formulations of the Problem

A natural idea is not always applicable. If for some pairs (i, j), we have both empirical correlations and expert estimates, then we can gauge the accuracy of expert estimates by comparing these values. But what if we do not have such pairs?

The main idea. The main idea is to use the fact that the actual correlation matrix a_{ij} must be non-negative definite, i.e., $\sum\limits_{i,j} a_{ij} \cdot z_i \cdot z_j \geq 0$ for all possible vectors $z = (z_1, \ldots, z_n)$.

Indeed, correlation a_{ij} between x_i and x_j is defined as

$$a_{ij} \stackrel{\text{def}}{=} \frac{E[\Delta x_i \cdot \Delta x_j]}{\sigma_i \cdot \sigma_j},$$

where $E[\cdot]$ means expected value, $\Delta x_i \stackrel{\text{def}}{=} x_i - E[x_i]$, and $\sigma_i^2 \stackrel{\text{def}}{=} E[(\Delta x_i)^2]$. Thus, the desired sum

$$S \stackrel{\text{def}}{=} \sum_{i,j} a_{ij} \cdot z_i \cdot z_j = \sum_{i,j} \frac{E[\Delta x_i \cdot \Delta x_j]}{\sigma_i \cdot \sigma_j} \cdot z_i \cdot z_j$$

can be equivalently expressed as

$$S = \sum_{i,j} E\left[\frac{z_i \cdot \Delta x_i}{\sigma_i} \cdot \frac{z_j \cdot \Delta x_j}{\sigma_j}\right],$$

i.e., as $S = E[s]$, where

$$s \stackrel{\text{def}}{=} \sum_{i,j} \frac{z_i \cdot \Delta x_i}{\sigma_i} \cdot \frac{z_j \cdot \Delta x_j}{\sigma_j}.$$

The expression s is nothing else but the square

$$s = \left(\sum_i \frac{z_i \cdot \Delta x_i}{\sigma_i}\right)^2.$$

This square s is always non-negative, thus its expected value $S = E[s]$ is also always non-negative, so the correlation matrix is indeed non-negative definite.

If the experts could provide the exact correlation values, then the matrix formed by their estimates would always be non-negative definite. However, as we have mentioned, the expert's estimates are approximate, i.e., they differ from the actual (unknown) correlation values. It is known that if we perturb a positive definite matrix by adding some random noise, then, if the noise is large enough, the perturbed matrix will stop being positive definite. If an expert provides many estimates, inevitably at some point, his estimates will violate the non-negative definiteness condition.

Our idea is thus to gauge the accuracy of the expert's estimates by computing how much we need to change the expert's estimates to make the matrix non-negative definite.

Let us describe this idea in precise terms. We will describe it for different settings.

Case when we only have experts' numerical estimates. In this case, we have a matrix $\widetilde{a}_{ij}$ formed by expert's estimates and 1s on the diagonal, and we want to find the smallest possible value $\varepsilon > 0$ for which there exists a non-negative definite matrix a_{ij} such that $a_{11} = \ldots = a_{nn} = 1$ and

$$|\widetilde{a}_{ij} - a_{ij}| \leq \varepsilon$$

for all $i \neq j$.

Once we find this value ε, we know that the expert's estimates may be ε-far from the actual correlation values. Thus, at best, based on each expert estimate $\widetilde{a}_{ij}$, we can only conclude that the actual value a_{ij} of the corresponding correlation is somewhere within the interval $[\widetilde{a}_{ij} - \varepsilon, \widetilde{a}_{ij} + \varepsilon]$.

Case when we have both some empirical correlations and expert's estimates of some correlations. In the previous case, we assumed that no empirical correlations are known at all, all correlation values come from an expert.

This may occur sometimes, but a more realistic situation is when we also know some empirical correlations. Empirical correlations are also approximate – any statistical estimates based on a finite sample are approximate – but these statistical correlations are usually much more accurate than the expert estimates, so in our analysis, we can safely ignore their inaccuracy and assume that these correlations are known exactly.

In this case:

- we know the exact values e_{ij} for some set of pairs S of different elements, and
- we know expert estimates $\widetilde{a}_{ij}$ for some set of pairs X of different elements.

In this case, we want to find the smallest possible value $\varepsilon > 0$ for which there exists a non-negative definite matrix a_{ij} such that:

- $a_{ii} = 1$ for all i,
- $a_{ij} = e_{ij}$ for all pairs $(i,j) \in S$, and
- $|\widetilde{a}_{ij} - a_{ij}| \le \varepsilon$ for all pairs $(i,j) \in X$.

Once we find this value ε, we can then conclude that a_{ij} belongs to the interval $[\widetilde{a}_{ij} - \varepsilon, \widetilde{a}_{ij} + \varepsilon]$.

Case when we have fuzzy estimates. For each pair $(i,j) \in X$ for which an expert provides an estimate, we can extract, in addition to the numerical estimate, a fuzzy number describing the expert's opinion about the correlation.

A fuzzy number describing each correlation a_{ij} can be alternatively described as a nested sequence of intervals $[a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$ (α-cuts) corresponding to different levels of confidence α.

Our goal is to find the largest α for which there exists a non-negative definite matrix a_{ij} such that:

- $a_{ii} = 1$ for all i,
- $a_{ij} = e_{ij}$ for all pairs $(i,j) \in S$, and
- $a_{ij} \in [a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$ for all pairs $(i,j) \in X$.

(We want the largest α, since the larger α, the narrower the intervals, and we want the narrowest intervals.)

Once we find this degree α, we can then conclude that for every i and j, the actual (unknown) value a_{ij} of the correlation belongs to the interval $[a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$.

III. What Is Known About This Problem and Similar Problems

What is known about this problem. We want to be able to check, for a given matrix $\widetilde{a}_{ij}$ and for a given $\varepsilon > 0$, whether in the ε-vicinity of this matrix there is a non-negative definite matrix. Unfortunately, in general, this problem is known to be NP-hard; see, e.g., [7].

This means, crudely speaking, that, unless P = NP (and most computer scientists believe that this is not possible), no feasible algorithm can exactly solve all particular cases of this problem.

This does not mean that we have to give up: many problems are NP-hard but become feasibly solvable when the approximation errors are relatively small – and this is the assumption we will make in this paper. If the approximation errors are huge, this means that experts are completely wrong, and, honestly, their estimates are practically useless.

What is known about similar problems. If instead of looking for the largest possible difference $\widetilde{a}_{ij} - a_{ij}$, we would look for the mean squared difference, then the corresponding problem becomes feasibly solvable; see, e.g., [1], [2], [3], [4], [5], [8], [10].

It is tempting to use the corresponding algorithms, but they are not good for our purpose: we are interested in individual values a_{ij}, and the fact that "on average" the deviation is small does not prevent us from the possibility that for this particular pair (i,j), the difference $\widetilde{a}_{ij} - a_{ij}$ is huge.

IV. Analysis of the Problem: Case of Numerical Expert Estimates

What we start with. For some pairs (i,j) – namely, for the pairs from the set S – we know, from experiments, the correlations e_{ij}. As we have mentioned, we can safely assume these correlations to be known exactly. For pairs that do not belong to the set S, all we know is the expert estimates $\widetilde{a}_{ij}$.

Thus, we get a matrix $a_{ij}^{(0)}$ whose elements are equal:

- to 1 when $i = j$,
- to e_{ij} when $(i,j) \in S$, and
- to $\widetilde{a}_{ij}$ when $(i,j) \notin S$.

In general, the resulting estimate for the correlation matrix may not be non-negative definite. As we have mentioned, due to the fact that expert estimates are approximate, the matrix $a_{ij}^{(0)}$ is only approximately equal to the actual correlation matrix, and may, thus, be not non-negative definite.

We would like to use this property to gauge the accuracy of expert estimates.

Reformulation in terms of eigenvalues. The above definition of a non-negative definite matrix requires that we try all possible vectors z. From the computational viewpoint, this is not realistic. However, there is a known easier-to-check equivalent property: namely, it is known that a matrix is non-negative definite if and only if all its eigenvalues are non-negative. This is easier to check, since there are efficient algorithms for computing the eigenvalues (and eigenvectors) of a matrix.

Thus, if a function is *not* non-negative definite, this means that some of the eigenvalues are negative.

The smallest of the negative eigenvalues is the most important one. When there is only one negative eigenvalue, then all we need is make it non-negative.

We want the smallest possible change, thus we want to select a non-negative eigenvalue which is the closest to the original negative eigenvalue. Such a value, of course, is 0. So, we want to change the original negative eigenvalue to 0.

We may have several negative eigenvalues. The smaller the negative eigenvalue, the more change we need to make to bring it to 0. Thus, to gauge the main effort, it is necessary to consider the smallest negative eigenvalue – i.e., the eigenvalue which is the furthest away from the desired set of all non-negative numbers.

How does the matrix change when we change eigenvalues.

- Let $\lambda < 0$ be the smallest negative eigenvalue of the matrix $a_{ij}^{(0)}$.
- Let us denote by $(e_1, \ldots, e_n)$ the corresponding unit eigenvector.

The fact this is a unit vector means that

$$\sum_{j=1}^{n} e_j^2 = 1,$$

and the fact that this vector is an eigenvector corresponding to the eigenvalue λ means that

$$\sum_j a_{ij}^{(0)} \cdot e_j = \lambda \cdot e_i.$$

The values $a_{ij}^{(0)}$ for $(i,j) \in S$ are known exactly, but the values corresponding to $(i,j) \notin S$ $(i \neq j)$ are known only approximately. We want to change the values $a_{ij}^{(0)}$ for all the pairs $(i,j) \notin S$ for which $i \neq j$, so that for the updated matrix $a_{ij}^{(0)} + \Delta a_{ij}$, the corresponding eigenvalue will be 0.

As we have mentioned, we consider the case when the estimation inaccuracy is relatively small, so that terms quadratic in terms of this inaccuracy can be safely ignored. Since the updated matrix $a_{ij}^{(0)} + \Delta a_{ij}$ is close to the original matrix $a_{ij}^{(0)}$, the corresponding unit eigenvector should be close to the original eigenvector e_j. Thus, the corresponding unit eigenvector of the updated matrix can be written as $e_j + \Delta e_j$, where the deviations Δe_j are small.

The fact that the vector $e_j + \Delta e_j$ is a unit vector means that

$$\sum_{j=1}^n (e_j + \Delta e_j)^2 = \sum_{j=1}^n \left(e_j^2 + 2e_j \cdot \Delta e_j + (\Delta e_j)^2\right) = 1,$$

i.e., that

$$\sum_{j=1}^n e_j^2 + 2\sum_{j=1}^n e_j \cdot \Delta e_j + \sum_{j=1}^n (\Delta e_j)^2 = 1.$$

This formula can be simplified due to the fact that:

- as we have mentioned, quadratic terms $(\Delta e_j)^2$ can be safely ignored, and
- the first sum is equal to 1:

$$\sum_{j=1}^n e_j^2 = 1.$$

Thus, the above condition takes the form

$$\sum_{j=1}^n e_j \cdot \Delta e_j = 0.$$

In geometric terms, this means that the deviation vector Δe_j is orthogonal to the original eigenvector e_j.

The condition that the new eigenvalue is 0 means that

$$\sum_{j=1}^n (a_{ij}^{(0)} + \Delta a_{ij}) \cdot (e_j + \Delta e_j) = 0$$

for all i. If we open the parentheses and ignore quadratic terms – i.e., for this formula, terms proportional to the product

$$\Delta a_{ij} \cdot \Delta e_j,$$

we get the following formula:

$$\sum_{j=1}^n a_{ij}^{(0)} \cdot e_j + \sum_{j=1}^n \Delta a_{ij} \cdot e_j + \sum_{j=1}^n a_{ij}^{(0)} \cdot \Delta e_j = 0.$$

The first term in the left-hand side is equal to $\lambda \cdot e_i$, so we conclude that

$$\sum_{j=1}^n a_{ij}^{(0)} \cdot \Delta e_j = |\lambda| \cdot e_i - \sum_{j=1}^n \Delta a_{ij} \cdot e_j.$$

This formula has the form

$$\sum_{j=1}^n a_{ij}^{(0)} \cdot \Delta e_j = f_i,$$

where we denoted

$$f_i \stackrel{\text{def}}{=} |\lambda| \cdot e_i - \sum_{j=1}^n \Delta a_{ij} \cdot e_j.$$

We have a system of linear equations for the unknowns Δe_j. We want to find a solution Δe_j of this system of linear equations, a solution which is orthogonal to the original eigenvector e_j.

The possibility of finding such a solution is the easiest to check if instead of the original orthonormal basis, we consider the othonormal basis consisting of eigenvectors of the original matrix $a_{ij}^{(0)}$.

- Let us denote the components of the vectors f_i and Δe_i in the new basis by F_k and ΔE_k, and
- let us denote the components of the matrix $a_{ij}^{(0)}$ in the new basis by $A_{k\ell}^{(0)}$.

In the new basis, the matrix $a_{ij}^{(0)}$ takes a diagonal form $A_{k\ell}^{(0)} = \lambda_k \cdot \delta_{k\ell}$, where λ_k is the k-th eigenvalue and $\delta_{k\ell}$ is the Kronecker symbol, i.e.:

- $\delta_{kk} = 1$ for all k, and
- $\delta_{k\ell} = 0$ for all $k \neq \ell$.

Thus, in the new basis, our system of linear equations takes the form $\lambda_k \cdot \Delta E_k = F_k$ for all k. The solution to this new system of equations is straightforward: we get a vector with components

$$\Delta E_k = \frac{F_k}{\lambda_k}.$$

Thus, in the original basis, the solution has the form

$$\Delta e_j = \sum_{k=1}^n \frac{F_k}{\lambda_k} \cdot e_j^{(k)},$$

where $e_j^{(k)}$ is the eigenvector corresponding to the k-th eigenvalue. In other words, the solution is a linear combination of different eigenvectors.

All eigenvectors from the original basis are orthogonal to each other. Thus, a linear combination of all eigenvectors different from the original eigenvector $e_j = e_j^{(k_0)}$ is also orthogonal to e_j. So, the requirement that this solution be orthogonal to the eigenvector e_j means that the corresponding component F_{k_0} should be 0.

For the orthonormal basis, this component is nothing else but a scalar (dot) product of the vector f_i and the unit

eigenvector e_i. Thus, for the equation to be solvable, this scalar product must be equal to 0:

$$\sum_{i=1}^{n} f_i \cdot e_i = \sum_{i=1}^{n} |\lambda| \cdot e_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta a_{ij} \cdot e_i \cdot e_j.$$

Here, since e_i is a unit vector, we have

$$\sum_{i=1}^{n} e_i^2 = 1.$$

The values Δa_{ij} are only different from 0 when $(i,j) \notin S$ and $i \neq j$. Thus, we must have

$$\sum_{(i,j)\notin S \,\&\, i\neq j} \Delta a_{ij} \cdot e_i \cdot e_j = |\lambda|. \tag{1}$$

Resulting reformulation of our problem: case of numerical estimates. In case of numerical estimates, the problem takes the following form: find the smallest possible value $\varepsilon > 0$ for which there exist values Δa_{ij} for which $|\Delta a_{ij}| \le \varepsilon$ for all i and j and for which the formula (1) is true.

How to solve the resulting optimization problem. For every ε, due to $|\Delta a_{ij}| \le \varepsilon$, we have $|\Delta a_{ij} \cdot e_i \cdot e_j| \le \varepsilon \cdot |e_i| \cdot |e_j|$. Thus,

$$\left|\sum_{(i,j)\notin S \,\&\, i\neq j} \Delta a_{ij} \cdot e_i \cdot e_j\right| \le \sum_{(i,j)\notin S \,\&\, i\neq j} |\Delta a_{ij} \cdot e_i \cdot e_j| \le$$

$$\sum_{(i,j)\notin S \,\&\, i\neq j} \varepsilon \cdot |e_i| \cdot |e_j| = \varepsilon \cdot S_0,$$

where we denoted

$$S_0 \stackrel{\text{def}}{=} \sum_{(i,j)\notin S \,\&\, i\neq j} |e_i| \cdot |e_j|.$$

So, if $\varepsilon \cdot S_0 < |\lambda|$, i.e., if

$$\varepsilon < \frac{|\lambda|}{S_0},$$

we cannot satisfy the formula (1).

When we reach the value

$$\varepsilon = \frac{|\lambda|}{S_0},$$

then it is already possible to satisfy the equation (1): namely, it is sufficient to take $\Delta a_{ij} = \varepsilon \cdot \mathrm{sign}(e_i) \cdot \mathrm{sign}(e_j)$, where $\mathrm{sign}(x)$ means the sign of the number x:

- $\mathrm{sign}(x) = 1$ when $x > 0$, and
- $\mathrm{sign}(x) = -1$ when $x < 0$.

For this choice of Δa_{ij}, we have

$$\sum_{(i,j)\notin S \,\&\, i\neq j} \Delta a_{ij} \cdot e_i \cdot e_j =$$

$$\varepsilon \cdot \sum_{(i,j)\notin S \,\&\, i\neq j} \mathrm{sign}(e_i) \cdot \mathrm{sign}(e_j) \cdot e_i \cdot e_j.$$

For every number x, we have $x \cdot \mathrm{sign}(x) = |x|$, thus, we get

$$\sum_{(i,j)\notin S \,\&\, i\neq j} \Delta a_{ij} \cdot e_i \cdot e_j = \varepsilon \cdot \sum_{(i,j)\notin S \,\&\, i\neq j} |e_i| \cdot |e_j| =$$

$$\varepsilon \cdot S_0 = \frac{|\lambda|}{S_0} \cdot S_0 = |\lambda|.$$

So, the smallest possible ε is equal to the ratio

$$\frac{|\lambda|}{S_0}.$$

What if for some pairs (i,j), we have both empirical correlations e_{ij} and expert estimates $\widetilde{a}_{ij}$? In this case, we need to take the difference between them into account as well. Thus, the smallest ε takes the form

$$\max\left(\frac{|\lambda|}{S_0}, \max_{(i,j)\in S\cap X} |\widetilde{a}_{ij} - e_{ij}|\right).$$

What if we only have expert estimates? In this case, the condition that $(i,j) \notin S$ and $i \neq j$ covers all the pairs – except for the pairs (i,i) for which the correlation is always 1 and hence, cannot be changed. Thus, here,

$$S_0 = \sum_{(i,j)\notin S \,\&\, i\neq j} |e_i| \cdot |e_j| = \sum_{i,j} |e_i| \cdot |e_j| - \sum_i |e_i|^2.$$

The first sum in the right-hand side is simply equal to the product

$$\left(\sum_{i=1}^{n} |e_i|\right) \cdot \left(\sum_{j=1}^{n} |e_j|\right),$$

i.e., to the square

$$\left(\sum_{i=1}^{n} |e_i|\right)^2.$$

Thus, we get

$$S_0 = \left(\sum_{i=1}^{n} |e_i|\right)^2 - \sum_{i=1}^{n} e_i^2.$$

So, we arrive at the following algorithm.

V. Resulting Algorithm

What is given.

- For some pairs (i,j), we are given the values e_{ij} of the empirical correlations. The set of all such pairs is denoted by S.
- For some pairs (i,j), we are given the expert estimates $\widetilde{a}_{ij}$ of the correlations. The set of all such pairs will be denoted by X.

We assume that for every pair of different indices, we have either an empirical value or an expert estimate (or both).

What we want to estimate. We want to estimate the accuracy of the expert estimates, i.e., the smallest ε for which we can

change the estimates $\widetilde{a}_{ij}$ by no more than ε and get a non-negative definite correlation matrix.

Cases. We consider two possible cases:

- the case when $S \cap X = \emptyset$, i.e., when for every pair, we have either an empirical correlation or an expert estimate, but not both; and
- the case when $S \cap X \neq \emptyset$, when for some pairs, we have both the empirical value of the correlation and the expert estimate.

In the first case, we will specifically consider the subcase when $S = \emptyset$, i.e., when we have no empirical correlation values, only expert estimates.

Algorithm.

1. Let us form a matrix $a_{ij}^{(0)}$ as follows:

- for $i = j$, we take $a_{ij}^{(0)} = 1$,
- for $(i, j) \in S$, we take $a_{ij}^{(0)} = e_{ij}$, and
- for $(i, j) \notin S$ and $i \neq j$, we take $a_{ij}^{(0)} = \widetilde{a}_{ij}$.

2. For the matrix $a_{ij}^{(0)}$, we compute the smallest eigenvalue λ. The following actions depend on whether this smallest eigenvalue is non-negative or negative.

2.1. If $\lambda \geq 0$, then the matrix $a_{ij}^{(0)}$ is already non-negative definite.

2.1.1. In the first case, when $S \cap X = \emptyset$, this means we cannot make any conclusions about the accuracy of the expert estimates: it could be that the expert estimates are exact.

2.1.2. In the second case, when $S \cap X \neq \emptyset$, as an estimate for expert accuracy, we take the largest difference between the expert estimates and the empirical correlations:

$$\varepsilon = \max_{(i,j) \in S \cap X} |\widetilde{a}_{ij} - e_{ij}|.$$

2.2. If $\lambda < 0$, then we compute the corresponding unit eigenvector e_i, and then we compute the value

$$S_0 = \sum_{(i,j) \notin S \,\&\, i \neq j} |e_i| \cdot |e_j|.$$

When $S = \emptyset$, we can compute S_0 by using a simplified formula

$$S_0 = \left(\sum_{i=1}^{n} |e_i| \right)^2 - \sum_{i=1}^{n} e_i^2.$$

The resulting estimate for ε depends on the case.

2.2.1. In the first case, when $S \cap X = \emptyset$, we take

$$\varepsilon = \frac{|\lambda|}{S_0}.$$

2.2.2. In the second case, when $S \cap X \neq \emptyset$, we take

$$\varepsilon = \max \left(\frac{|\lambda|}{S_0}, \max_{(i,j) \in S \cap X} |\widetilde{a}_{ij} - e_{ij}| \right).$$

VI. Fuzzy Case: Analysis of the Problem

What is the problem: reminder. Which α should we select?

- First, we want to make sure that when for some pairs, we have both empirical correlations and expert estimates, the empirical correlation lies within the corresponding interval, i.e., that for all such pairs, we have

$$a_{ij}^{-}(\alpha) \leq e_{ij} \leq a_{ij}^{+}(\alpha).$$

- Second, we want to make sure that within selected intervals, we have values a_{ij} for which the correlation matrix is non-negative definite.

We have nested intervals. The intervals $[a_{ij}^{-}(\alpha), a_{ij}^{+}(\alpha)]$ grow when α decreases. Thus, if the above two conditions are satisfied for some α, they are also satisfied for all smaller values $\alpha' < \alpha$ as well.

Bisection idea. Thus, to find the largest α for which both conditions are satisfied, we can use the following natural *bisection* idea:

- First, we check whether both conditions are satisfied for $\alpha = 1$. If they are satisfied, then $\alpha = 1$ is the value we take.
- If one or both of the above conditions are not satisfied for $\alpha = 1$, then we check whether they are satisfied for

$$\alpha = 0.$$

- If they are not even satisfied for $\alpha = 0$, this means that the expert underestimates his/her uncertainty, so we cannot reply on this fuzzy information to gauge this uncertainty.
- If both conditions are satisfied for $\alpha = 0$, this means that we have:
  - a value $\underline{\alpha}$ (in this case, $\underline{\alpha} = 0$) for which both conditions are satisfied, and
  - a value $\overline{\alpha}$ (in this case, $\overline{\alpha} = 1$) for which at least one of the conditions is not satisfied.
- In this case, we know that the desired value α is somewhere between $\underline{\alpha}$ and $\overline{\alpha}$, i.e., somewhere on the interval

$$[\underline{\alpha}, \overline{\alpha}].$$

- Once we get this information, we can check whether both conditions are satisfied for the midpoint

$$\alpha_m = \frac{\underline{\alpha} + \overline{\alpha}}{2}$$

of this interval.
  - If both conditions are satisfied, then we have a new interval $[\alpha_m, \overline{\alpha}]$ of half the size that contains the desired value α.
  - On the other hand, if at least one of the conditions is not satisfied, then we also have a new interval of half size containing α: namely, the interval $[\underline{\alpha}, \alpha_m]$.
- In both cases, we decrease the size of the interval in half.

How many iterations do we need?

- We start with an interval $[0, 1]$ of width 1.

- Thus, in four iterations, we get an interval of width $1/2^4 = 1/16 = 0.0625$.
- Experts do not describe their degree of certainty with higher accuracy than one decimal digit.
- So, 4 iterations are more than enough for finding the main digit of the desired value α.

Remaining question. The remaining question is how, given α, we can check whether both conditions are satisfied.

- Checking the first condition is easy: we simply check the corresponding inequalities.
- How can we check the second condition?

To answer this question, let us recall the above case – when we had numerical estimates.

A fuzzy estimate is an extension of a numerical estimate. A fuzzy estimate for the correlation a_{ij} is an extension of the numerical estimate. We start with the numerical value – which corresponds to the degree of certainty 1, and we add intervals containing this value: the smaller the degree of confidence that all these values are indeed possible, the wider the interval. In this case, the numerical value corresponds to the top α-cut, corresponding to $\alpha = 1$: $\widetilde{a}_{ij} = a_{ij}^-(1) = a_{ij}^+(1)$.

Sometimes, experts start not with a numerical estimate, but with an interval $[a_{ij}^-(1), a_{ij}^+(1)]$ of positive width. In this case, it makes sense to take, as a representative numerical value, the midpoint of this interval

$$\widetilde{a}_{ij} = \frac{a_{ij}^-(1) + a_{ij}^+(1)}{2}.$$

How to check non-negative definiteness. For each α, possible values of a_{ij} lie within the interval $[a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$. Thus, possible values of $\Delta a_{ij} = a_{ij} - \widetilde{a}_{ij}$ lie between $a_{ij}^-(\alpha) - \widetilde{a}_{ij}$ and $a_{ij}^+(\alpha) - \widetilde{a}_{ij}$.

Non-negative positiveness, as we have shown, means that we must have

$$\sum_{(i,j) \notin S \,\&\, i \neq j} \Delta a_{ij} \cdot e_i \cdot e_j \geq |\lambda|$$

for some Δa_{ij}.

What is the largest value that the sum in the left-hand side of this inequality can take? To find out, let us describe the largest possible value of each of the terms $\Delta a_{ij} \cdot e_i \cdot e_j$.

- When the product $e_i \cdot e_j$ is positive, then the maximum of this term is attained for positive values Δa_{ij}. The largest positive value v_{ij} of Δa_{ij} is equal to $a_{ij}^+(\alpha) - \widetilde{a}_{ij}$.
- When the product $e_i \cdot e_j$ is negative, the maximum of the i-th term in the sum is attained for negative values Δa_{ij}. The largest absolute value of these negative values is $v_{ij} = \widetilde{a}_{ij} - a_{ij}^-(\alpha)$.

In both cases, for each term $\Delta a_{ij} \cdot e_i \cdot e_j$, the largest possible value of this term is $v_{ij} \cdot |e_i| \cdot |e_j|$. Thus, the largest possible value of the desired sum is equal to

$$\sum_{(i,j) \notin S \,\&\, i \neq j} v_{ij} \cdot |e_i| \cdot |e_j|.$$

- If this sum is smaller than $|\lambda|$, this means we cannot reach $|\lambda|$ by selecting appropriate deviations – and thus, that the corresponding value α is too large.
- On the other hand, if this sum is larger than or equal to $|\lambda|$, this means that for this α, it is possible to attain non-negative positiveness.

So, we arrive at the following algorithm.

VII. Fuzzy Case: Algorithm

What is given.

- For some pairs (i, j), we know the empirical correlations e_{ij}. The set of all such pairs will be denoted by S.
- For some pairs (i, j), experts give us fuzzy estimates $[a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$ corresponding to different values α. The set of all such pairs (i, j) will be denoted by X.

We assume that for every pair of different indices, we have either an empirical value or an expert estimate (or both).

What is our objective. Our goal is to return the value α so that, for each $(i, j) \notin S$ for which $i \neq j$, the interval $[a_{ij}^-(\alpha), a_{ij}^+(\alpha)]$ is used as the range of possible values of correlation.

Algorithm: preliminary stage.

- First, for each $(i, j) \notin S$ for which $i \neq j$, we compute the value
$$\widetilde{a}_{ij} = \frac{a_{ij}^-(1) + a_{ij}^+(1)}{2}.$$
- We then compute the following matrix $a_{ij}^{(0)}$:
  - for $i = j$, we take $a_{ij}^{(0)} = 1$,
  - for $(i, j) \in S$, we take $a_{ij}^{(0)} = e_{ij}$, and
  - for $(i, j) \notin S$ and $i \neq j$, we take $a_{ij}^{(0)} = \widetilde{a}_{ij}$.
- After that, we compute the smallest eigenvalue λ of the matrix $a_{ij}^{(0)}$ and the corresponding unit eigenvector e_i.

The results of all these preliminary computations are used in the main stage of the algorithm.

Algorithm: main stage. Once the first stage is over, the main stage starts.

- First, we use an auxiliary algorithm – described below – to check whether the above-mentioned conditions are satisfied for $\alpha = 1$. If they are satisfied, we return $\alpha = 1$ and stop.
- If the conditions are not satisfied for $\alpha = 1$, we check whether they are satisfied for $\alpha = 0$. If they are not satisfied, then we ignore all the fuzzy information as useless and use only the numerical values $\widetilde{a}_{ij}$ as described in the previous section.
- If the conditions are satisfied for $\alpha = 0$ and not satisfied for $\alpha = 1$, then we set $\underline{\alpha} = 0$ and $\overline{\alpha} = 1$ and start iterations.
- On each iteration, we check whether the condition is satisfied for
$$\alpha_m = \frac{\underline{\alpha} + \overline{\alpha}}{2}.$$

- If the conditions are satisfied for α_m, then we replace $\underline{\alpha}$ with α_m, while keeping $\overline{\alpha}$ unchanged.
- If the conditions are not satisfied for α_m, then we replace $\overline{\alpha}$ with α_m while keeping $\underline{\alpha}$ unchanged.

- Iterations stop when $\overline{\alpha} - \underline{\alpha} \le \delta$ for a given δ (e.g., for $\delta = 0.1$). At this point, we return the midpoint

$$\alpha_m = \frac{\underline{\alpha} + \overline{\alpha}}{2}$$

as the desired value α.

Auxiliary algorithm. In this algorithm, we are also given a number α, and we want to check whether the conditions are satisfied for this α.

- First, we check whether for all $(i, j) \in S \cap X$, we have

$$a_{ij}^-(\alpha) \le e_{ij} \le a_{ij}^+(\alpha).$$

If at least one of these inequalities is not satisfied, we stop the auxiliary algorithm and conclude that the conditions are not satisfied for this α.
- If all the above inequalities are satisfied and $\lambda \ge 0$, we conclude that both conditions are satisfied.
- If all double inequalities are satisfied but $\lambda < 0$, then for all $(i, j) \notin S$ for which $i \ne j$, we compute the following value v_{ij}:
  - when $\mathrm{sign}(e_i) \cdot \mathrm{sign}(e_j) > 0$, we take

$$v_{ij} = a_{ij}^+(\alpha) - \widetilde{a}_{ij};$$

  - when $\mathrm{sign}(e_i) \cdot \mathrm{sign}(e_j) < 0$, we take

$$v_{ij} = \widetilde{a}_{ij} - a_{ij}^-(\alpha).$$

- Then, we check whether

$$\sum_{(i,j) \notin S \,\&\, i \ne j} v_{ij} \cdot |e_i| \cdot |e_j| \ge |\lambda|.$$

  - If this inequality is satisfied, we conclude that both conditions are satisfied for the given α.
  - If this inequality is not satisfied, we conclude that at least one of the conditions is *not* satisfied for the given α.

VIII. Numerical Example

Let us illustrate our algorithm on a simple numerical expert estimate. We assume that in a variability quantification the following correlation matrix $a_{ij}^{(0)}$ has been derived from expert estimates:

$$\begin{bmatrix} 1 & 0.6 & -0.6 \\ 0.6 & 1 & 0.6 \\ -0.6 & 0.6 & 1 \end{bmatrix}.$$

The eigenvalues λ_k are as follows:

$$\begin{bmatrix} -0.2 \\ 1.6 \\ 1.6 \end{bmatrix}.$$

Because there is one negative eigenvalue $\lambda = -0.2$, the correlation matrix is not non-negative definite.

The corresponding eigenvector e_i is:

$$\begin{bmatrix} -\dfrac{1}{\sqrt{3}} \\ \dfrac{1}{\sqrt{3}} \\ -\dfrac{1}{\sqrt{3}} \end{bmatrix}.$$

Using the provided algorithms to generate the closest non-negative definite correlation matrix, the smallest possible change for all elements of a_{ij} is

$$\varepsilon = \frac{|\lambda|}{S_0} = \frac{|\lambda|}{\left(\sum\limits_{i=1}^{n} |e_i|\right)^2 - \sum\limits_{i=1}^{n} e_i^2} = 0.1.$$

Via the changes

$$\Delta a_{ij} = \varepsilon \cdot \mathrm{sign}(e_i) \cdot \mathrm{sign}(e_j),$$

the mathematically valid non-negative definite correlation matrix is then obtained:

$$a_{ij} = \begin{bmatrix} 1 & 0.5 & -0.5 \\ 0.5 & 1 & 0.5 \\ -0.5 & 0.5 & 1 \end{bmatrix}.$$

Now the eigenvalues are:

$$\begin{bmatrix} 0 \\ 1.5 \\ 1.5 \end{bmatrix}.$$

All eigenvalues are equal to or larger than 0, hence we double-checked that the correlation matrix is non-negative definite after the conversion, with the largest diversion of the expert's estimates $\varepsilon = 0.1$.

IX. Conclusions

In this paper, the uncertainty of the expert-estimated correlation is quantified, in a way of satisfying the mathematical meaning of a valid correlation matrix—being non-negative definite. We provide algorithms to generate a valid correlation matrix out of an invalid one. For the fuzzy situation, we provide algorithms to give the narrowest interval (with the given confidence) containing a non-negative definite correlation matrix.

Acknowledgments

This work was performed when Zitong Gong and Vladik Kreinovich were visiting researchers at the Leibniz Universität Hannover, supported by the German Research Foundation (DFG), partially under a Mercator Fellowship within the Research Training Group GRK2159 (i.c.sens) and partially under the research project D5 "Risk Assessment of Regeneration Paths for Supporting Simultaneous Decisions" within the Collaborative Research Center (CRC) 871 – Regeneration of Complex Capital Goods.

This work was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation, and by a program of the China Scholarship Council.

REFERENCES

[1] R. Borsdorf, *A Newton Algorithm for the Nearest Correlation Matrix*, M.Sc. Thesis, University of Manchester, 2007.

[2] R. Borsdorf, *Structured Matrix Nearness Problems: Theory and Applications*, PhD Dissertation, University of Manchester, 2012.

[3] R. Borsdorf and N. J. Higham, "A preconditioned Newton algorithm for the nearest correlation matrix", *IMA Journal of Numerical Analysis*, 2010, Vol. 30, pp. 94–107.

[4] R. Borsdorf, N. J. Higham, and M. Raydan, "Computing a nearest correlation matrix with factor structure", *SIAM Journal on Matrix Analysis and Applications*, 2010, Vol. 31, pp. 2603–2622.

[5] N. J. Higham, "Computing the nearest correlation matrix – a problem from finance", *IMA Journal of Numerical Analysis*, 2002, Vol. 22, pp. 329–343.

[6] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[7] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

[8] C. Lucas, *Computing Nearest Covariance and Correlation Matrices*, M.Sc. Thesis, University of Manchester, 2001.

[9] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[10] Huo-Duo Qi and Defeng Sun, "A quadratically confergent Newton method for computing the nearest correlation matrix", *SIAM Journal on Matrix Analysis and Applications*, 2006, Vol. 28, pp. 360–385.

[11] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.