What If We Do Not Know Correlations?

Michael Beer[1,3], Zitong Gong[3], Ingo Neumann[2],
Songsak Sriboonchitta[4], and Vladik Kreinovich[5]

[1]Institute for Risk and Reliability and [2]Geodetic Institute
Leibniz University Hannover, 30167 Hannover, Germany
beer@irz.uni-hannover.de, neumann@gih.uni-hannover.de
[3]Institute for Risk and Uncertainty, University of Liverpool
Liverpool L69 3BX, UK, Zitong.Gong@liverpool.ac.uk
[4]Faculty of Economics, Chiang Mai University, Thailand
songsakecon@gmail.com
[5]University of Texas at El Paso, USA, vladik@utep.edu

1. Need for Data Processing

- In many real-life situations, we are interesting in quantities y which are difficult to measure directly.
- For example, we may be interested in the distance to a faraway star or in the amount of oil in a given oil field.
- Since we cannot measure y directly, a natural idea is to measure it *indirectly*, i.e.,
  - to find easier-to-measure quantities $x_1, \ldots, x_n$ which are connected to y by a known algorithm

$$y = f(x_1, \ldots, x_n),$$

  - and use the results $\widetilde{x}_i$ of measuring x_i to estimate y as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

2. What is the Accuracy of the Resulting Estimate?

- The results $\widetilde{x}_i$ of measuring x_i are, in general, different from the actual values of the measured quantities.
- In other words, there is a usually a measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, so that $x_i = \widetilde{x}_i - \Delta x_i$.
- As a result, the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is also, in general, different from the actual value

$$y = f(x_1, \ldots, x_n) = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n).$$

- It is therefore desirable to estimate the error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of the indirect measurement.

3. Measurement Errors are Usually Relatively Small

- In most real-life situations, the measurement errors are relatively small.
- As a result, we can safely ignore terms which are quadratic (or of higher order) with respect to Δx_i.
- For example, if the measurement error is 10%, its square is 1%, which is much smaller.
- We know that

$$\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n).$$

- So, we can expand this expression in Taylor series in Δx_i and keep only linear terms in this expansion.
- As a result, $\Delta y = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$, where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}_{|(\widetilde{x}_1, \ldots, \widetilde{x}_n)}$.

4. What Do We Know About Δx_i

- In the ideal case, for each measuring instrument, we know the first two moments of the measurement errors:
  - we know the mean value μ_i of the corresponding measurement error Δx_i, and
  - we know the standard deviation σ_i.
- If we know the exact mean, then:
  - we can re-calibrate the i-th measuring instrument by subtracting μ_i from all the measurement results;
  - in this case, we get the mean value equal to 0.
- Often, we only know the mean and the standard deviation with uncertainty, i.e., we only know that

$$\underline{\mu}_i \le \mu_i \le \overline{\mu}_i \text{ and } \underline{\sigma}_i \le \sigma_i \le \overline{\sigma}_i.$$

5. Based on This Information, We Can Estimate the Mean Value μ of Δy

- Based on this information, we can estimate the mean μ of the desired measurement error.
- Namely, it follows that $\mu = \sum\limits_{i=1}^{n} c_i \cdot \mu_i$.
- So, if we know the exact values of means μ_i, we can use this formula to find μ.
- If μ_i are only known with interval uncertainty, then we can represent the interval $[\underline{\mu}_i, \overline{\mu}_i]$ in the centered form

$$[\widetilde{\mu}_i - \Delta_i, \widetilde{\mu}_i + \Delta_i], \text{ where } \widetilde{\mu}_i \stackrel{\text{def}}{=} \frac{\underline{\mu}_i + \overline{\mu}_i}{2}, \quad \Delta_i \stackrel{\text{def}}{=} \frac{\overline{\mu}_i - \underline{\mu}_i}{2}.$$

- Then, each $\mu_i \in [\underline{\mu}_i, \overline{\mu}_i] = [\widetilde{\mu}_i - \Delta_i, \widetilde{\mu}_i + \Delta_i]$ can be represented as $\widetilde{\mu}_i + \Delta\mu_i$, where $\Delta\mu_i \stackrel{\text{def}}{=} \mu_i - \widetilde{\mu}_i \in [-\Delta_i, \Delta_i]$.

6. Estimating μ (cont-d)

- Then, $\mu = \widetilde{\mu} + \Delta\mu : \widetilde{\mu} \stackrel{\text{def}}{=} \sum_{i=1}^{n} c_i \cdot \widetilde{\mu}_i$, $\Delta\mu \stackrel{\text{def}}{=} \sum_{i=1}^{n} c_i \cdot \Delta\mu_i$.
- The largest value of $\Delta\mu$ is attained when each of the terms $c_i \cdot \Delta\mu_i$ is the largest.
- For $c_i > 0$, this happens when $\Delta\mu_i$ is the largest, i.e., when $\Delta\mu_i = \Delta_i$.
- For $c_i \le 0$, this happens when $\Delta\mu_i$ is the smallest, i.e., when $\Delta\mu_i = -\Delta_i$.
- In both cases, the largest value of $c_i \cdot \Delta\mu_i$ is $|c_i| \cdot \Delta_i$.
- Similarly, the smallest value of $c_i \cdot \Delta\mu_i$ is $-|c_i| \cdot \Delta_i$.
- Thus, $\mu \in [\widetilde{\mu} - \Delta, \widetilde{\mu} + \Delta]$, where $\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \Delta_i$.

7. What is the Standard Deviation σ of Δy: Case When We Know the Correlations

- To complete our description of the uncertainty Δy, we need to also estimate its standard deviation σ.
- This is equivalent to estimating the variance $V = \sigma^2 = E[(\delta y)^2]$, where $\delta y \stackrel{\text{def}}{=} \Delta y - E[\Delta y] = \Delta y - \mu$.
- Here, $\delta y = \sum_{i=1}^{n} c_i \cdot \delta x_i$, where $\delta x_i \stackrel{\text{def}}{=} \Delta x_i - E[\Delta x_i] = \Delta x_i - \mu_i$.
- Thus, $E[(\delta y)^2] = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i \cdot c_j \cdot E[\delta x_i \cdot \delta x_j]$.
- For $i = j$, we get $E[(\delta x_i)^2] = \sigma_i^2$.
- For $i \neq j$, by definition of the correlation ρ_{ij}, we have $\rho_{ij} = \dfrac{E[\delta x_i \cdot \delta x_j]}{\sigma_i \cdot \sigma_j}$, thus $E[\delta x_i \cdot \delta x_j] = \rho_{ij} \cdot \sigma_i \cdot \sigma_j$.

8. Estimating σ: Case When We Know the Correlations (cont-d)

- We know that $\sigma^2 == \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} c_i \cdot c_j \cdot E[\delta x_i \cdot \delta x_j]$, where $E[(\delta x_i)^2] = \sigma_i^2$ and $E[\delta x_i \cdot \delta x_j] = \rho_{ij} \cdot \sigma_i \cdot \sigma_j$.
- So, $\sigma^2 = \sum\limits_{i=1}^{n} c_i^2 \cdot \sigma_i^2 + \sum\limits_{i \neq j} \rho_{ij} \cdot c_i \cdot c_j \cdot \sigma_i \cdot \sigma_j$.
- So, if we know ρ_{ij}, we can estimate the desired standard deviation σ of the result y of data processing.

9. But What If We Do Not Know the Correlations?

- In some practical situations, however, we do not know the correlations.
- In this case, depending on the actual values of the correlations, we get different values σ.
- What is the range of possible values σ?
- This is the question that we answer in this talk.

10. First Result

- We consider the case when:
  - we know the exact values of the standard deviations σ_i, but
  - we have no information about the correlations.
- Then, the range of possible values of σ is $[\underline{\sigma}, \overline{\sigma}]$, where

$$\overline{\sigma} = \sum_{i=1}^{n} |c_i| \cdot \sigma_i, \quad \underline{\sigma} = \max\left(0, |c_{i_0}| \cdot \sigma_{i_0} - \sum_{i \neq i_0}^{n} |c_i| \cdot \sigma_i\right), \text{ and}$$

$$i_0 \text{ is the index for which } |c_{i_0}| \cdot \sigma_{i_0} = \max_i |c_i| \cdot \sigma_i.$$

- *Comment:* the formula for $\overline{\sigma}$ is, surprisingly, very similar to the formula for $\overline{\mu}$.

11. General Result

- Now, we assume that we only know the bounds $\underline{\sigma}_i$ and $\overline{\sigma}_i$ on the standard deviations.
- We still assume that we have no information about correlations.
- Then the range of possible values of σ is $[\underline{\sigma}, \overline{\sigma}]$, where

$$\overline{\sigma} = \sum_{i=1}^{n} |c_i| \cdot \overline{\sigma}_i, \quad \underline{\sigma} = \max\left(0, |c_{i_0}| \cdot \underline{\sigma}_{i_0} - \sum_{i \neq i_0} |c_i| \cdot \overline{\sigma}_i\right).$$

- Here, i_0 is the index for which the product $|c_{i_0}| \cdot \underline{\sigma}_{i_0}$ is the largest.
- If there are several such indices i_0, then we select the one for which the product $|c_{i_0}| \cdot \overline{\sigma}_{i_0}$ is the smallest.

12. Acknowledgments

- We acknowledge support of Center of Excellence in Econometrics, Chiang Mai University, Thailand.
- This work was performed
  - when Vladik was a visiting researcher with the Geodetic Institute, Leibniz University of Hannover,
  - this visit was supported by the German Science Foundation.
- This work was also supported in part by the US National Science Foundation grant HRD-1242122.

13. Proof of First Result

- It is well known that for every two random variables a and b, we have

$$\sigma^2[a+b] = \sigma^2[a] + \sigma^2[b] + \rho_{ab} \cdot \sigma[a] \cdot \sigma[b].$$

- Since the correlation coefficient ρ_{ab} is always bounded by 1, we conclude that

$$\sigma^2[a+b] \leq \sigma^2[a] + \sigma^2[b] + 2\sigma[a] \cdot \sigma[b].$$

- The right-hand side of this inequality is $(\sigma[a]+\sigma[b])^2$, thus we conclude that

$$\sigma[a+b] \leq \sigma[a] + \sigma[b].$$

- In particular, for $a-b$ and b, we thus get

$$\sigma[a] \leq \sigma[a-b] + \sigma[b], \text{ hence } \sigma[a-b] \geq \sigma[a] - \sigma[b].$$

- Let us apply these inequalities to our case.

14. Proof: Part 2

- The overall random component $\delta y = \Delta y - E[\Delta y]$ of the measurement error Δy is the sum of n terms $c_i \cdot \delta x_i$.
- For each term $c_i \cdot \delta x_i$, the standard deviation is $|c_i| \cdot \sigma_i$.
- Thus, the st. dev. σ of the sum δy of these terms does not exceed the sum of st. dev.:

$$\sigma \le \sum_{i=1}^{n} |c_i| \cdot \sigma_i.$$

- Alternatively, we can represent δy as the difference $\delta y = c_{i_0} \cdot \delta x_{i_0} - s$, where $s \stackrel{\text{def}}{=} \sum\limits_{i \ne i_0} (-c_i) \cdot \delta x_i$.
- Thus, by using the formula for the standard deviation of the difference, we get $\sigma \ge |c_{i_0}| \cdot \sigma[s]$.
- By using the formula for the standard deviation of the sum, we conclude that $\sigma[s] \le \sum\limits_{i \ne i_0} |c_i| \cdot \sigma_i$.

15. Proof: Part 2 (cont-d)

- Thus, we have $\sigma \ge |c_{i_0}| \cdot \sigma_{i_0} - \sum\limits_{i \ne i_0} |c_i| \cdot \sigma_i$.
- Clearly also $\sigma \ge 0$, so

$$\sigma \ge \max\left(|c_{i_0}| \cdot \sigma_{i_0} - \sum_{i \ne i_0} |c_i| \cdot \sigma_i\right).$$

- So, we proved that for the above expressions for $\underline{\sigma}$ and $\overline{\sigma}$, we always have $\underline{\sigma} \le \sigma \le \overline{\sigma}$.
- To complete our proof, it is now sufficient to prove that the values $\underline{\sigma}$ and $\overline{\sigma}$ are attainable.

16. Proof: Part 3

- Let us first prove that the upper bound $\overline{\sigma}$ is attainable.
- Indeed, let η be a standard normally distributed random variable, with 0 mean and standard deviation 1.
- Then, we can take $\delta x_i = \text{sign}(c_i) \cdot \sigma_i \cdot \eta$, where

$$\text{sign}(x) \stackrel{\text{def}}{=} 1 \text{ for } x \geq 0 \text{ and } \text{sign}(x) \stackrel{\text{def}}{=} -1 \text{ for } x < 0.$$

- Then, $\text{sign}(x) \cdot x = |x|$ for all x, so:

$$\delta y = \sum_{i=1}^{n} c_i \cdot \delta_i = \sum_{i=1}^{n} c_i \cdot \text{sign}(c_i) \cdot \sigma_i \cdot \eta = \sum_{i=1}^{n} |c_i| \cdot \sigma_i \cdot \eta = \left(\sum_{i=1}^{n} |c_i| \cdot \sigma_i \right) \cdot \eta.$$

- This sum has the desired standard deviation $\sum\limits_{i=1}^{n} |c_i| \cdot \sigma_i$.

17. Proof: Part 4

- Let's prove that the lower bound is also attainable.
- We will first prove it for the case when the difference $d \stackrel{\text{def}}{=} |c_{i_0}| \cdot \sigma_{i_0} - \sum\limits_{i \ne i_0} |c_i| \cdot \sigma_i$ is positive; then, $\underline{\sigma} = d$.
- Take $\delta x_{i_0} = \text{sign}(c_{i_0}) \cdot \sigma_{i_0} \cdot \eta$, $\delta x_i = -\text{sign}(c_i) \cdot \sigma_i \cdot \eta$ for all $i \ne i_0$; then:

$$\delta y = c_{i_0} \cdot \delta x_{i_0} + \sum_{i \ne i_0} c_i \cdot \delta x_i = |c_{i_0}| \cdot \sigma_{i_0} \cdot \eta - \sum_{i \ne i_0} |c_i| \cdot \sigma_i \cdot \eta =$$

$$\left(|c_{i_0}| \cdot \sigma_{i_0} \eta - \sum_{i \ne i_0} |c_i| \cdot \sigma_i \right) \cdot \eta = d \cdot \eta.$$

- Since $d > 0$, this sum has standard deviation $d = \underline{\sigma}$.

18. Proof: Part 5

- To finalize the proof, we need to show that when $d < 0$, the sum Δy can have zero standard deviation.
- Let us prove, by induction over m, the following auxiliary result: when $a_1 \le \ldots \le a_m$, then:
  - then for every number a from $\max\left(0, a_m - \sum_{i=1}^{m-1} a_i\right)$ to $\sum_{i=1}^{m} a_i$,
  - there exist planar vectors A_i for which $|A_i| = a_i$ for all i and $\left|\sum_{i=1}^{m} A_i\right| = a$.
- The base case $m = 2$ is straightforward.
- Indeed, in this case, the desired inequality takes the form $a_2 - a_1 \le a \le a_2 + a_1$.

19. Proof: Part 5 (cont-d)

- To get a vector A with $|A| = a_1 + a_2$, we simply take A_1 and A_2 parallel and going in the same direction.
- To get a vector A with $|A| = a_2 - a_1$, we take A_1 and A_2 parallel but going in different directions.
- By a continuous transformation of one configuration into another, we get cases with all intermediate a's.
- Let us now describe the induction step.
- Suppose that we have already proved this result for m, we want to prove it for $m + 1$.
- The value $a = a_1 + \ldots + a_m + a_{m+1}$ is easy to obtain: take A_i parallel and going in the same direction.

20. Proof: Part 5 (cont-d)

- If $a_{m+1} > a_1 + \ldots + a_m$, then the value $a = a_{m+1} - \sum\limits_{i=1}^{m} a_i$ is also easy to obtain:
  - we take all the vector parallel,
  - the first m vectors $A_1, \ldots, A_m$ go in one direction, and
  - the vector A_{m+1} goes in the opposite direction.
- To complete the proof of induction step, we need to consider the case when $a_{m+1} < a_1 + \ldots + a_m$.
- In this case, we want to find the vectors for which the sum is 0.
- By induction assumption:
  - for the sum $A_1 + \ldots + A_m$,
  - any length from $\max(0, a_m - (a_1 + \ldots + a_{m-1}))$ to $a_1 + \ldots + a_m$ is possible.

21. Proof: Part 5 (cont-d)

- Here, $a_{m+1} < a_1 + \ldots + a_m$, since this is the case that we are considering.
- Also, $a_{m+1} \geq 0$ and $a_{m+1} \geq a_m$ hence $a_{m+1} \geq a_m - \sum\limits_{i=1}^{m-1} a_i$ and thus $a_{m+1} \geq \max\left(0, a_m - \sum\limits_{i=1}^{m-1} a_i\right)$.
- So, by induction assumption, there exist vectors $A_1, \ldots, A_m$ for which $|A_1 + \ldots + A_m| = a_{m+1}$.
- Now, if we take $A_{m+1} = -(A_1 + \ldots + A_m)$, we get $|A_{m+1}| = a_{m+1}$ and $A_1 + \ldots + A_m + A_{m+1} = 0$.
- The auxiliary statement is proven.

22. Proof: Part 5 (cont-d)

- The auxiliary statement implies that:
  - when a_{i_0} is larger than or equal to all the values a_i and $a_{i_0} \le \sum\limits_{i \ne i_0} a_i$,
  - then there exist planar vectors A_i of lengths $|A_i| = a_i$ for which $\sum\limits_i A_i = 0$.
- Let us take such A_i corr. to $a_i = |c_i| \cdot \sigma_i$; let us:
  - select two independent normal random variables η' and η'', with 0 mean and st. dev. 1, and
  - assign, to each planar vector A with coordinates $A = (A', A'')$, a random variable $\eta_A \stackrel{\text{def}}{=} A' \cdot \eta' + A'' \cdot \eta''$.
- One can easily check that $V[\eta_A] = (A')^2 + (A'')^2 = |A|^2$, where $|A|$ is the length of A.

23. Proof: Part 5 (cont-d)

- Thus, $\sigma[\eta_A] = |A|$.
- It is also easy to check that the transformation $A \to \eta_A$ from vectors to random variables is linear:

$$\eta_{c_A \cdot A + \ldots + c_B \cdot B} = c_A \cdot \eta_A + \ldots + c_B \cdot \eta_B.$$

- We can then take for each i, as δx_i, the random variable corresponding to the vector $\dfrac{A_i}{c_i}$.
- This variable has standard deviation

$$\left|\frac{A_i}{c_i}\right| = \frac{|A_i|}{|c_i|} = \frac{|c_i| \cdot \sigma_i}{|c_i|} = \sigma_i.$$

- Here, $c_i \cdot \delta x_i = \eta_{A_i}$.

24. Proof: Part 5 (final)

- We have shown that $c_i \cdot \delta x_i = \eta_{A_i}$.
- Thus, for the sum $\delta y = \sum\limits_{i=1}^{n} c_i \cdot \delta x_i$, we have

$$\delta y = \sum_{i=1}^{n} c_i \cdot \delta x_i = \sum_{i=1}^{n} \eta_{A_i} = \eta_{\sum\limits_{i=1}^{n} A_i} = \eta_0 = 0.$$

- The statement is proven, and so is our first result.

25. Proof of the General Result

- This proof if straightforward.
- For example, for the upper bound,
  - from the fact that for all possible values σ_i, we get $\sigma \le \sum_{i=1}^{n} |c_i| \cdot \sigma_i$ and that $\sigma_i \le \overline{\sigma}_i$,
  - we conclude that $\sigma \le \sum_{i=1}^{n} |c_i| \cdot \overline{\sigma}_i$.
- Vice versa:
  - by taking $\sigma_i = \overline{\sigma}_i$ in the example from the proof of the previous result,
  - we get an example when σ is equal to the upper bound $\sum_{i=1}^{n} |c_i| \cdot \sigma_i$.
- To get a similar example for the lower bound, we should take $\sigma_{i_0} = \underline{\sigma}_{i_0}$ and $\sigma_i = \overline{\sigma}_i$ for all $i \ne i_0$.