How to Describe Correlation in the Interval Case?

Carlos Jimenez, Francisco Zapata, and Vladik Kreinovich

University of Texas at El Paso, El Paso, TX 79968, USA
cjimenez23@miners.utep.edu, fazg74@gmail.com, vladik@utep.edu
1. Correlations Are Ubiquitous

- One of the main objectives of science and engineering is to improve the world,
  - to enhance good things and
  - to make sure that bad things do not happen.
- The state of the world is usually described by the values of different quantities.
- In these terms, our objective is to change the values of the corresponding quantities:
  - to increase the economy’s growth rate,
  - to decrease unemployment,
  - to decrease the patient’s body temperature or blood pressure, etc.
- In many practical situations, we cannot change these quantities directly.
2. Correlations Are Ubiquitous (cont-d)

• Thus, the only way to change them is to change them indirectly: i.e.,
  – to find auxiliary possible-to-change quantities that are correlated with the desired ones
  – in the sense that changes in these auxiliary quantities will lead to the desired changes in the quantities of interest.

• For example:
  – a change in the central bank’s interest rate or a change in tax rules can boost the economy,
  – a change in the patient’s diet and/or exercise schedule can lower his/her blood pressure, etc.

• In some cases, we know which two quantities are correlated.
3. Correlations Are Ubiquitous (cont-d)

- In many other situations, we are actively looking for quantities which are correlated with the desired ones.

- For example:
  - for many diseases,
  - we are actively looking for ways to control the genes that would help fight these diseases.

- Looking for correlations is important.

- It is therefore important to have an adequate description of this intuitive notion.
4. What Is Correlation: Main Idea

- The main idea: the use of \( x \) can improve our ability to predict \( y \).

- In other words, correlation means that
  - if we take \( x \) into account,
  - then we can get more accurate predictions of \( y \) than if we don’t.

- Similarly, the absence of correlation means that the use of \( x \) cannot help in predicting \( y \).

- For example, intuitively, fluctuations of a quasar’s flux are not related to weather.

- This means that:
  - even if we add quasar’s flux as a possible additional variable into the weather prediction models,
  - we will not get more accurate predictions.
5. What Is Correlation (cont-d)

- To describe this idea in precise terms, we need to formally describe:
  - what models we consider and
  - how we measure model’s accuracy.
6. Need for Linear Models

- When we do not consider $x$ at all, then the only possible models for $y$ are models in which $y = \text{const.}$
- When we take $x$ into account, we thus get models of the type $y = f(x)$, for some function $f(x)$.
- Which functions should we consider?
- In most cases, changes in both $x$ and $y$ are small.
- We are happy when the growth rate increases from 2% to 3%.
- We are happy when the upper blood pressure falls from 140 to 130, etc.
7. **Need for Linear Models (cont-d)**

- When changes in $x$ are small, i.e., when all the values $x$ have the form $x_0 + \Delta x$ for some small $\Delta x$, then:
  
  - we can expand the dependence $f(x) = f(x_0 + \Delta x)$ on $\Delta x$ and
  
  - ignore terms which are quadratic or of higher order in terms of $\Delta x$.

- In this case, we get a linear model $f(x) = a_0 + a_1 \cdot \Delta x$.

- Substituting $\Delta x = x - x_0$ into this expression, we conclude that

  $$f(x) = a_0 + a_1 \cdot (x - x_0) = (a_0 - a_1 \cdot x_0) + a_1 \cdot x.$$ 

- Thus, it makes sense to restrict ourselves to linear models.
8. How to Gauge Accuracy: Traditional Approach

- Models are practically always approximate.
- It is very rare to have a model that enables us to predict the exact value of a quantity.
- Many independent reasons cause model’s predictions $f(x_i)$ to be different from the actual values $y_i$.
- Thus, the difference $\Delta y_i = y_i - f(x_i)$ is the sum of many independent random variables.
- Most of these variables are of about the same size.
9. How to Gauge Accuracy (cont-d)

- In probability theory, there is a result – known as Central Limit Theorem – according to which,
  - when the number of components is large,
  - the distribution of the sum of many small independent components is close to Gaussian (normal).
- The larger the number of such components, the closer we are to a Gaussian distribution.
- In practice, we usually have many different reasons causing the model to differ from reality.
- So, we can safely assume that the difference $\Delta y_i$ is normally distributed.
- A normal distribution for $\Delta y$ can be characterized by its mean $\mu$ and its standard deviation $\sigma$. 
10. How to Gauge Accuracy (cont-d)

• Different reasons cause lead to positive and negative differences; so, it is reasonable to assume that:
  – on average, such reasons cancel each other and
  – the mean values of the difference is 0.

• So, the only parameter that describes the model’s accuracy is the standard deviation $\sigma$.

• Factors influencing different measurements are, in general, independent.

• Thus, the differences $\Delta y_i$ corresponding to different measurements $i$ are independent.

• Since the mean is 0, the square $\sigma^2$ of the standard deviation – i.e., the variance – can be estimated as

$$\sigma^2 \approx \frac{1}{n} \cdot \sum_{i=1}^{n} (\Delta y_i)^2.$$
11. How to Gauge Accuracy (cont-d)

- Here \( n \) denotes the overall number of measurements.
- We want to find the most accurate model, i.e., the model for which \( \sigma \) is as small as possible.
- Minimizing \( \sigma \) is equivalent to minimizing \( \sigma^2 \), which, in its turn, is equivalent to minimizing the sum \( \sum_{i=1}^{n} (\Delta y_i)^2 \).
- Here, we are minimizing the sum of the squares (of differences).
- So, this method of finding the most adequate model is known as the Least Squares Method.
12. Resulting Formula for Correlation

- If we do not take $x$ into account, then the only models we have are the models $y = \text{const}$. 

- To find the best such model, we find the constant for which the corresponding variance is the smallest:

$$\sigma_y^2 = \min_a \left( \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - a)^2 \right).$$

- If we take $x$ into account, then we allow models of the type $y \approx a + b \cdot x$.

- Then, for the best such model, we get the variance

$$\sigma_{y|x}^2 = \min_{a,b} \left( \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2 \right).$$

- If $x$ and $y$ are not correlated, then the use of $x$ will lead to more accurate models for $y$: $\sigma_{y|x}^2 = \sigma_y^2$. 
13. Resulting Formula for Correlation (cont-d)

- On the other hand, if $y$ is uniquely determined by $x$, i.e., if $y = a + b \cdot x$, then $\sigma^2_{y|x} = 0 \ll \sigma_y^2$.

- In general, intuitively, the larger part of original variance is decreased by using $x$, the larger the correlation.

- So, it’s reasonable to define correlation as

$$C_{y|x} = 1 - \frac{\sigma^2_{y|x}}{\sigma_y^2}.$$

- This intuitive idea is well described by the usual statistical correlation: $C_{y|x} = \rho^2_{xy}$, where

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \cdot \sigma_y}, \quad C_{xy} \overset{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}),$$

$$\bar{x} \overset{\text{def}}{=} \frac{1}{n} \cdot x_i, \quad \bar{y} \overset{\text{def}}{=} \frac{1}{n} \cdot y_i, \quad \sigma^2_x \overset{\text{def}}{=} \frac{1}{n} \cdot (x_i - \bar{x})^2, \quad \text{and} \quad \sigma^2_y \overset{\text{def}}{=} \frac{1}{n} \cdot (y_i - \bar{y})^2.$$
14. Need to Go Beyond Normal Distributions

- We assumed that many independent factors are of approximately the same size.
- Then the differences $\Delta y = y - f(x)$ are normally distributed.
- In practice, however, there may be a few major reasons for the difference.
- In this case, the quantity $\Delta y$ is not necessarily normally distributed.
- In this case, what is the reasonable formalization of the intuitive notion of correlation?
15. How to Gauge Model Accuracy

• We do not know the probability distribution of the model inaccuracy $\Delta y$.

• So, a natural idea is to consider the absolute values of this inaccuracy; namely:
  
  – if for one model, we always have $|\Delta y| \leq \Delta_1$,
  
  – and for another model, we always have $|\Delta y| \leq \Delta_2$ with $\Delta_2 < \Delta_1$,
  
  – then the second model is more accurate than the first one.

• As a measure of model’s accuracy, it is therefore reasonable to take the smallest $\Delta$ for which $|\Delta y| \leq \Delta$:

\[
\Delta = \max_i |\Delta y_i| = \max_i |y_i - f(x_i)|.
\]
16. Relation to Interval Uncertainty

- For all $x$, we have $|\Delta y| = |y - f(x)| \leq \Delta$.
- This means that for each $x$, the value $y$ belongs to the interval $[f(x) - \Delta, f(x) + \Delta]$.
- Thus, the above case corresponds to interval uncertainty.
17. Resulting Definition of Correlation

- If we do not use $x$, then the only possible models are constant models $y = b$.

- The accuracy of the best such model can be described by the quantity $\Delta_y = \min_a \left( \max_i |y_i - a| \right)$.

- One can easily check that the corresponding value $a$ is equal to $a = \frac{1}{2} \cdot \left( \min_i y_i + \max_i y_i \right)$.

- The corresponding value $\Delta_y = \frac{1}{2} \cdot \left( \max_i y_i - \min_i y_i \right)$.

- If we allow $x$, then the best accuracy of the corresponding linear models $y \approx a + b \cdot x$ is

  $\Delta_{y|x} = \min_{a,b} \left( \max_i |y_i - (a + b \cdot x_i)| \right)$.

- Similarly to the usual case, it is therefore reasonable to define correlation as $\rho_{y|x}^\text{int} = 1 - \frac{\Delta_{y|x}}{\Delta_y}$. 
18. Open Question

- The usual statistical correlation is symmetric:
  \[ \rho_{xy} = \rho_{yx}. \]

- Is the interval analogue of correlation symmetric?
19. Case of Non-Linear Dependence

- The actual dependence is sometimes non-linear.
- It is thus reasonable to also include, e.g., quadratic (or even cubic) terms in the corresponding model.
- Then we consider, e.g., the values

\[
\sigma^2_{y|x} = \min_{a,b,c} \left( \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - (a + b \cdot x_i + c \cdot x_i^2)|^2 \right) \quad \text{or} \quad \\
\Delta y|x = \min_{a,b,c} \left( \max_i |y_i - (a + b \cdot x_i + c \cdot x_i^2)| \right).
\]

- In addition to such quadratic etc. polynomials, we can also consider other families of models.
20. Dependence on Several Variables

- We can consider dependence on different quantities $x_1, \ldots, x_k$, e.g., as $C_{y|x_1,\ldots,x_k}^{\text{int}} = 1 - \frac{\sigma_y^2}{\sigma_y^2 |x_1,\ldots,x_k|}$, with $\sigma_y^2 |x_1,\ldots,x_k| \stackrel{\text{def}}{=} \min_{a,b_1,\ldots,b_k} \left( \frac{1}{n} \cdot \sum_i \left| y_i - (a + b_1 \cdot x_{1i} + \ldots + b_k \cdot x_{ki}) \right|^2 \right)$. 

- We can also take $C_{y|x_1,\ldots,x_k}^{\text{int}} = 1 - \frac{\Delta_y |x_1,\ldots,x_k|}{\Delta_y}$, where

$$\Delta_y |x_1,\ldots,x_k| = \min_{a,b_1,\ldots,b_k} \left( \max_i \left| y_i - (a + b_1 \cdot x_{1i} + \ldots + b_k \cdot x_{ki}) \right| \right).$$
21. Robust Techniques

- We can also consider cases of robust statistics – when we do not know the probability distribution.
- It is known as robust statistics.
- An example are $\ell^p$-methods in which the model’s accuracy is described by a value $s$ for which

$$s^p = \frac{1}{n} \cdot \sum_{i=1}^{n} |\Delta y_i|^p.$$  

- Then, we can define $C_{p,y|x} = 1 - \frac{s^p_{y|x}}{s^p_y}$, where

$$s^p_{y|x} = \min_a \left( \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - a|^p \right) \quad \text{and} \quad s^p_y = \min_a \left( \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - a|^p \right)$$

$$s^p_{y|x} = \min_{a,b} \left( \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - (a + b \cdot x_i)|^p \right).$$
22. Acknowledgments

This work was supported in part by the National Science Foundation grant HRD-1242122 (Cyber-ShARE Center).