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

Abstract

In many areas of science and engineering, we want to change a difficult-to-directly-change quantity – e.g., the economy's growth rate. Since we cannot directly change the desired quantity, we need to find easier-to-change auxiliary quantities that are correlated with the desired quantity – in the sense that a change in the auxiliary quantity will cause the change in the desired quantity as well. How can we describe this intuitive notion of correlation in precise terms? The traditional notion of correlation comes from situations in which there are many independent factors causing the predictive model to differ from the actual values and all these factors are of about the same size. In this case, the distribution of the difference between the model's predictions and the actual values is close to normal. In many practical situations, however, there are a few major factors which are much larger than others. In this case, the distribution of the differences is not necessarily normal. In this paper, we show how, in such situations, we can formalize the intuitive notion of correlation.

1 Formulation of the Problem: Need to Extend the Intuitive Notion of Correlation Beyond Normal Distributions

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. 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. However, in many other situations – e.g., in many medical research projects – 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.

In view of importance of looking for correlation, it is important to have an adequate description of this intuitive notion.

What is correlation: main idea. The main idea behind the intuitive notion of correlation between the quantities x and y is that the use of x can improve our ability to predict y. In other words, correlation means that if we take x into account, 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.

To describe this idea in precise terms, we need to formally describe what models we consider and how we measure model's accuracy.

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

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. Usually, there are many different independent reasons why the model's predictions are different from the actual value of the corresponding quantity. Thus, the difference between the model's prediction and the actual value is the sum of many independent random variables – most of which are of about the same size.

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) distribution – and the larger the number of such components, the closer we are to a Gaussian distribution; see, e.g., [7]. Since in practice, we usually have many different reasons causing the model to differ from reality, we can safely assume that the difference between the model's predictions and the actual value is normally distributed.

A normal distribution for $\Delta y \stackrel{\text{def}}{=} y - f(x)$ can be characterized by its mean μ and its standard deviation σ. 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 σ.

Factors influencing different measurements are, in general, independent. Therefore, it is reasonable to conclude that the differences $\Delta y_i = y_i - f(x_i)$ corresponding to different measurements i are independent random variables.

Since the mean is 0, the square σ^2 of the standard deviation – i.e., the variance – can be estimated as the mean value of $(\Delta y_i)^2$, i.e., as $\dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (\Delta y_i)^2$, where n denotes the overall number of measurements.

We want to find the most accurate model, i.e., the model for which the standard deviation σ is as small as possible. Minimizing σ is equivalent to minimizing σ^2, which, in its turn, is equivalent to minimizing the sum $\sum\limits_{i=1}^{n} (\Delta y_i)^2$. This method of finding the most adequate model is known as the *Least Squares Method*, since we are minimizing the sum of the squares (of differences); see, e.g., [7].

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, i.e., if 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 – i.e., we will have $\sigma^2_{y|x} = \sigma^2_y$. On the other hand, if y is uniquely determined by x, i.e., if $y = a + b \cdot x$, then $\sigma^2_{y|x} = 0 \ll \sigma^2_y$. In general, intuitively, the larger part of original variance is decreased by using x, the larger the correlation. So, it is reasonable to define correlation as

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

It turns out that this intuitive idea is well described by the usual statistical correlation: namely, $C_{y|x} = \rho^2_{xy}$, where

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

$$\overline{x} \stackrel{\text{def}}{=} \frac{1}{n} \cdot x_i, \quad \overline{y} \stackrel{\text{def}}{=} \frac{1}{n} \cdot y_i, \quad \sigma^2_x \stackrel{\text{def}}{=} \frac{1}{n} \cdot (x_i - \overline{x})^2, \text{ and } \sigma^2_y \stackrel{\text{def}}{=} \frac{1}{n} \cdot (y_i - \overline{y})^2.$$

Need to go beyond normal distributions. When we have many independent factors causing the model $f(x)$ to deviate from the actual values y, and all the 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 Δy is not necessarily normally distributed. In this case, what is the reasonable formalization of the intuitive notion of correlation?

2 How to Describe Correlation in the Interval Case: Main Idea

How to gauge model accuracy. In situations when we do not know the probability distribution of the model inaccuracy Δy, a natural idea is to consider the absolute values of this inaccuracy. Namely, if:

- for one model, we always have $|\Delta y| \le \Delta_1$,
- for another model, we always have $|\Delta y| \le \Delta_2$,
- and $\Delta_2 < \Delta_1$,

this means that 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 possible value Δ for which $|\Delta y| \le \Delta$ – i.e., the value

$$\Delta = \max_i |\Delta y_i| = \max_i |y_i - f(x_i)|.$$

Relation to interval uncertainty. If for all x, we have

$$|\Delta y| = |y - f(x)| \le \Delta,$$

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*; see, e.g., [2, 4, 5, 6].

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

$$a = \frac{1}{2} \cdot \left(\min_i y_i + \max_i y_i \right),$$

and the corresponding value Δ_y is equal to

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

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

$$\rho^{\rm int}_{y|x} = 1 - \frac{\Delta_{y|x}}{\Delta_y}.$$

Open question. The usual statistical correlation is symmetric: $\rho_{xy} = \rho_{yx}$. Is the interval analogue of correlation symmetric?

3 Additional Ideas

Case of non-linear dependence. If the actual dependence is non-linear, it is reasonable to also include, e.g., quadratic (or even cubic) terms in the corresponding model, and consider, e.g., the values

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

or

$$\Delta_{y|x} = \min_{a,b,c} \left(\max_i \left| y_i - \left(a + b \cdot x_i + c \cdot x_i^2 \right) \right| \right);$$

see, e.g., [3]. In addition to such quadratic etc. polynomials, we can also consider other families of models.

Dependence on several variables. We can also consider dependence on different quantities $x_1, \ldots, x_k$, e.g., as

$$C^{\rm int}_{y|x_1,\ldots,x_k} = 1 - \frac{\sigma^2_{y|x_1,\ldots,x_k}}{\sigma^2_y},$$

where

$$\sigma^2_{y|x_1,\ldots,x_k} = \min_{a,b_1,\ldots,b_k} \left(\frac{1}{n} \cdot \sum_i |y_i - (a + b_1 \cdot x_{1i} + \ldots + b_k \cdot x_{ki})|^2 \right),$$

or

$$C^{\rm int}_{y|x_1,\ldots,x_k} = 1 - \frac{\Delta_{y|x_1,\ldots,x_k}}{\Delta_y},$$

where

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

Robust techniques. In addition to normal distributions and interval uncertainty, we can also consider cases of statistics developed for situation when we do not know the probability distribution – so-called robust statistics (see, e.g., [1]): for example, ℓ^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

$$s^p_y = \min_a \left(\frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - a|^p \right),$$

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

and

$$C_{p,y|x} = 1 - \frac{s^p_{y|x}}{s^p_y}.$$

Acknowledgments

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

References

[1] P. J. Huber and E. M. Ronchetti, Robust Statistics, Wiley, Hoboken, New Jersey, 2009.

[2] L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2001.

[3] V. Kreinovich, H. T. Nguyen, and S. Sriboonchitta, "How to Detect Linear Dependence on the Copula Level?", In: V.-N. Huynh, V. Kreinovich, and S. Sriboonchitta (eds.), *Modeling Dependence in Econometrics*, Springer Verlag, Berlin, Heidelberg, 2014, pp. 65–82.

[4] G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.

[5] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[6] S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, Berlin, 2005.

[7] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.