How to Distinguish True Dependence from Varying Independence?

Marketa Krmelova[1], Martin Trnecka[1],
Vladik Kreinovich[2], and Berlin Wu[3]
[1]Department of Computer Science
Palacky University
Olomouc 17. listopadu 12
CZ-771 46 Olomouc, Czech Republic
marketa.krmelova@gmail.com
martin.trnecka@gmail.com
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu
[3]Department of Mathematical Sciences
National Chengchi University
Taipei 116, Taiwan
berlin@nccu.edu.tw

Abstract

A usual statistical criterion for the quantities X and Y to be independent is that the corresponding distribution function $F(x, y)$ is equal to the product of the corresponding marginal distribution functions. If this equality is violated, this is usually taken to mean that X and Y are dependent. In practice, however, the inequality may be caused by the fact that we have a mixture of several populations, in each of which X and Y are independent. In this paper, we show how we can distinguish true dependence from such varying independence. This can also lead to new measures to degree of independence and of varying independence.

1 Formulation of the Problem

Independence: a usual description (see, e.g., [8])**.** In statistics, independence between two events A and B means that the probability of the event A is not affected by whether the event B occurs or not. For example, the prob-

ability of winning a lottery is the same where a person was born in January or not. In precise terms, this means that the conditional probability $P(A \,|\, B)$ of A under the condition B is equal to the probability $P(A)$ of the event A: $P(A \,|\, B) = P(A)$.

Since $P(A \,|\, B) = \dfrac{P(A \,\&\, B)}{P(B)}$, the above equality is equivalent to

$$P(A \,\&\, B) = P(A) \cdot P(B).$$

For quantities X and Y, independence means that every event related to X is independent from any event related to Y. In particular, for every two real numbers x and y, the events $X \le x$ and $Y \le y$ are independent. In precise terms, this means that

$$F(x, y) = F_X(x) \cdot F_Y(y), \tag{1}$$

where $F(x, y) \stackrel{\text{def}}{=} P(X \le x \,\&\, Y \le y)$ is a joint (cumulative) distribution function of the pair (X, Y) , and $F_X(x) \stackrel{\text{def}}{=} P(X \le x)$ and $F_Y(y) \stackrel{\text{def}}{=} P(Y \le y)$ are marginal distribution functions.

To derive the formula (1), we used very specific events $X \le x$ and $Y \le y$. However, one can show that once the formula (1) is satisfied, each event related to X is independent from each event related to Y. Thus, the formula (1) can be used as the definition of independence.

Alternatively, independence can be described in terms of the probability density functions of the corresponding distributions:

$$\rho(x, y) = \rho_X(x) \cdot \rho_Y(y), \tag{2}$$

where

$$\rho(x, y) \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{P(x - h \le X \le x + h \,\&\, y - h \le Y \le y + h)}{(2h) \cdot (2h)}$$

is the probability density of the joint distribution, and

$$\rho_X(x) = \lim_{h \to 0} \frac{P(x - h \le X \le x + h)}{2h} \text{ and } \rho_Y(y) = \lim_{h \to 0} \frac{P(y - h \le Y \le y + h)}{2h}$$

are probability densities of the marginal distributions.

Varying independence: idea. In many real-life situations, we have a mixture of several populations for each of which X and Y are independent. For example, suppose that we analyze the dependence of salary on height, and suppose that within each country, salary and height are independent. However, if we bring together some people from Sweden (where people are taller and salaries are higher) and from Greece (where people are somewhat shorter and salaries are somewhat smaller), we may get a wrong conclusion that salary and height are related.

This is an example of what we call *varying* independence: for each population, X and Y are independent, but because we have a mixture of populations, we get an illusion of dependence.

Varying independence: description in precise terms. Let K be a total number of different populations, let w_k ($1 \le k \le K$) denote the probability that a randomly selected object belongs to the k-th population, and let $A_k(x)$ and $B_k(y)$ be marginal distribution functions corresponding to the k-th population.

We assumed that within each population, the quantities X and Y are independent. Thus, for each population k, the joint probability distribution has the form $A_k(x) \cdot B_k(y)$, and the overall probability distribution has the form

$$F(x,y) = \sum_{k=1}^{K} w_k \cdot A_k(x) \cdot B_k(y). \tag{3}$$

Similar description in terms of probability densities. A similar formula describes the *probability density* function $\rho(x,y)$ of the joint distribution (3):

$$\rho(x,y) = \sum_{k=1}^{K} w_k \cdot a_k(x) \cdot b_k(y), \tag{4}$$

where $a_k(x)$ and $b_k(y)$ are the probabilities densities of the marginal distributions corresponding to the k-th population.

Problem: how to distinguish true dependence from varying independence? A natural question is: how can we separate the situations when we have two truly dependent variables X and Y from situations of varying independence, for some small number of populations K?

This is the question that we will be answering in this paper.

2 Idealized Situation

Description of the idealized situation. To answer the above question, let us start with an ideal situation, in which we know the exact values of the probability distribution $F(x,y)$ for all x and y, or, to be more precise, we know the exact values $F(x_i, y_j)$ for all the values (x_i, y_j) for some dense grid $x_i = x_0 + i \cdot h_x$ ($1 \le i \le I$) and $y_j = y_0 + j \cdot h_y$ ($1 \le j \le J$).

Analysis of the ideal situation. In this case, for each i, the vector

$$\vec{F}_i \stackrel{\text{def}}{=} (F(x_i, y_1), \ldots, F(x_i, y_J))$$

is a linear combination of K vectors $\vec{B}_k \stackrel{\text{def}}{=} (B_k(y_1), \ldots, B_k(y_J))$:

$$\vec{F}_i = \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot \vec{B}_k. \tag{5}$$

Thus, every $K+1$ vectors of this type will be linearly dependent.

Vice versa, if every $K+1$ vectors $\vec{F}_i$ are linearly dependent, this means that we can represent all the vectors $\vec{F}_i$ as a linear combination of $\le K$ basic vectors, i.e., that we have an expression (5) for all i. The equality between two vectors means that all their components are equal, so we get (3) for all x and y.

This analysis leads us to the following way of checking whether a given function $F(x,y)$ has a representation of type (3).

Resulting algorithm. We are given a function $F(x,y)$ and an integer K, and we want to check whether the given function can be represented in the form (3). For this checking, we consider vectors $\vec{F}_i \stackrel{\text{def}}{=} (F(x_i,y_1),\ldots,F(x_i,y_J))$ one by one.

We try $i'=1,2,\ldots$, and out of the first i' vectors we find the set S consisting of K linearly independent ones. We start with an empty set S.

- Once the set S is formed for some i', we check whether the next vector $\vec{F}_{i'+1}$ is linearly independent from S (see, e.g., [1]).
- If the vector $\vec{F}_{i'+1}$ is a linear combination of vectors from the set S, then we keep the set S intact and go to the next value i'.
- If $\vec{F}_{i'+1}$ is linearly independent from S, then we add this vector $\vec{F}_{i'+1}$ to the set S, and also go to the next value i'.

If at some point, we get a set S with $>K$ elements, we stop and conclude that a representation of type (3) is impossible, i.e., we have a true dependence. On the other hand, if after considering all I vectors $\vec{F}_1$, $\ldots$, $\vec{F}_I$, we have a set S with $\le K$ vectors, this means that a representation of type (3) is possible, i.e., we have a varying independence.

Comment. Instead of the values $F(x_i,y_j)$ of the joint distribution function, we can similarly consider the values $\rho(x_i,y_j)$ of the joint probability density function.

3 General Case

Description of the general case. In general, we only know approximate values of the distribution $F(x,y)$. In this case, instead of the exact equality (3), we have an approximate equality

$$F(x_i,y_j) \approx \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot B_k(y_j). \tag{6}$$

Formal description of the problem. A usual statistics-motivated way to deal with approximate equalities is to use the least squares approach, i.e., to look for the the values w_k and the functions $A_k(x_i)$ and $B_k(y_j)$ for which the sum of the least squares

$$s \stackrel{\text{def}}{=} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(F(x_i, y_j) - \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot B_k(y_j) \right)^2 \tag{7}$$

attains the smallest possible value.

If, for this minimum $s_{\min}$, the corresponding average value $a \stackrel{\text{def}}{=} \dfrac{s_{\min}}{I \cdot J}$ does not exceed the accuracy σ^2 with which we know the values $F(x_i, y_j)$, this means that the measurements are consistent with the varying independence. On the other hand, if $a > \sigma^2$, this means that varying independence cannot explain the observed data, i.e., we have true dependence.

Analysis of the problem. We want to find K vectors $\vec{B}_k$ so that all I vectors $\vec{F}_i$ can be approximately represented as linear combinations of these vectors. For each i, the corresponding sum

$$\sum_{j=1}^{J} \left(F(x_i, y_j) - \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot B_k(y_j) \right)^2 \tag{8}$$

is the square of the Euclidean (l_2) distance between the vector $\vec{F}_i$ and the corresponding linear combination, and

$$s_i \stackrel{\text{def}}{=} \min_{w_k, A_k(x_i)} \sum_{j=1}^{J} \left(F(x_i, y_j) - \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot B_k(y_j) \right)^2 \tag{9}$$

is the square of the smallest such distance, i.e., the square

$$d^2(\vec{F}_i, \mathrm{Lin}(\vec{B}_1, \ldots, \vec{B}_K))$$

of the distance between the vector $\vec{F}_i$ and the K-dimensional linear space generated by the vectors $\vec{B}_k$ ($1 \le k \le K$). In these terms, we are given I vectors $\vec{F}_i$, and we need to find the K-dimensional linear space for which the sum of the distance $s_1 + \ldots + s_I$ is the smallest possible.

It is known (see, e.g., [2, 4]) that the solution to this problem is the linear space spanned by the first K *singular vectors* $\vec{v}_1, \ldots, \vec{v}_K$ of the matrix $F(x_i, y_j)$, i.e., the unit vectors for which

$$\vec{v}_1 = \arg \max_{\|\vec{v}\|=1} \|F\vec{v}\|, \tag{10}$$

where F is the matrix with elements $F(x_i, y_j)$, $\|\vec{v}\| \stackrel{\text{def}}{=} \sqrt{v_1^2 + \ldots + v_J^2}$ is the usual Euclidean norm of a vector $\vec{v} = (v_1, \ldots, v_J)$, and for every $k > 1$,

$$\vec{v}_k = \arg \max_{\|\vec{v}\|=1, \vec{v} \perp \vec{v}_1, \ldots, \vec{v} \perp \vec{v}_{k-1}} \|F\vec{v}\|. \tag{11}$$

The actual minimum $s_{\min}$ can be described in terms of the corresponding singular values $\sigma_k \stackrel{\text{def}}{=} \|F\vec{v}_k\|$, as

$$s_{\min} = \|F\|_F^2 - \sum_{k=1}^{K} \sigma_k^2, \tag{12}$$

where

$$\|F\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}(F(x_i, y_j))^2} \tag{13}$$

is the *Frobenius norm* of the matrix F. There are many efficient algorithm for *Singular Value Decomposition* (SVD), i.e., for computing the corresponding singular vectors and singular values.

Comment. It is worth mentioning that squares of singular values are eigenvalues of the matrices $F^T F$ and FF^T, i.e., matrices with elements

$$M_{ii'} \stackrel{\text{def}}{=} \sum_{j=1}^{J} F(x_i, y_j) \cdot F(x_{i'}, y_j) \text{ and } M'_{jj'} \stackrel{\text{def}}{=} \sum_{i=1}^{I} F(x_i, y_j) \cdot F(x_i, y_{j'}).$$

Resulting algorithm. The above analysis leads to the following algorithm for checking whether a given joint distribution $F(x_i, y_j)$ corresponds to varying dependence with a given integer K and accuracy σ. For this checking:

- We first compute the Frobenius norm (13) of the matrix F with the components $F(x_i, y_j)$.
- Then, we apply SVD to the matrix F to compute the first K singular values $\sigma_1, \ldots, \sigma_K$.
- After that, we compute $s_{\min}$ by using the formula (12).
- Finally, we compute $a = \dfrac{s_{\min}}{I \cdot J}$.
  - If $a \le \sigma^2$, then the available information is consistent with the varying dependence.
  - If $a > \sigma^2$, then we have true dependence.

Comments.

- If we are interested in the corresponding probability distributions $A_k(x)$ and $B_k(y)$, then we can find them as linear combinations of the corresponding singular vectors $\vec{v}_k$.
- Alternatively, we can apply a similar analysis to the matrix ρ whose components are the values $\rho(x_i, y_j)$ of the probability density.

- A similar problem can be formulated – and the same solution proposed – when we have more than two variables X, Y, ..., Z. The main challenge here is that while for two quantities X and Y, there exist efficient algorithms for solving the corresponding problem, for several variables, the corresponding decomposition problem is proven to be computationally intractable (NP-hard); see, e.g., [3, 6]. Since in this case, there is no universally applicable feasible algorithm, we have to use heuristic methods; see, e.g., a survey [6].

4 Related Measures of Dependence and Varying Dependence

Formulation of the problem. When the two quantities are not exactly independent and are not exactly consistent with the assumption of varying independence, it is reasonable to ask how close the resulting distribution is to independence or to varying independence.

Possible solution: idea. In view of the above analysis, a reasonable solution to this problem is to use the smallest l^2-distance between the given distribution and possible K-varying independent ones as the desired measure. This leads us to the following definition.

Definition 1. *Let the values $F(x_i, y_j)$ of the probability distribution be given, and let a positive integer K be given. By a* measure of deviation from K-varying independence*, we mean the value $d_K \stackrel{\text{def}}{=} \sqrt{\frac{s_{\min}}{I \cdot J}}$, where $s_{\min}$ is the smallest possible value of the quantity*

$$s \stackrel{\text{def}}{=} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(F(x_i, y_j) - \sum_{k=1}^{K} w_k \cdot A_k(x_i) \cdot B_k(y_j) \right)^2 \tag{7}$$

over all possible values of w_k, $A_k(x_i)$, and $B_k(y_j)$.

Comments.

- For $K = 1$, we get a measure of deviation from independence, i.e., a measure of dependence.
- We can get alternative measures of dependence if instead of using the values $F(x_i, y_j)$ of the probability distribution, we use the values $\rho(x_i, y_j)$ of the probability density function.

How to efficiently compute the resulting measures of dependence. Similarly to the previous section, we can use known efficient algorithms for SVD to compute the corresponding dependence measures d_K. Namely:

- We first compute the Frobenius norm (13) of the matrix F with the components $F(x_i, y_j)$.
- Then, we apply SVD to the matrix F to compute the first K singular values $\sigma_1, \ldots, \sigma_K$.
- After that, we compute $s_{\min}$ by using the formula (12).
- Finally, we compute $d_K = \sqrt{\frac{s_{\min}}{I \cdot J}}$.

5 Formulations in Terms of Copulas

Formulation of the problem. Whether two quantities X and Y are dependent or independent does not change if we apply linear or non-linear transformations $X \to X' = f(X)$ and $Y \to Y' = g(Y)$ to each of these quantities. However, for the above definition of dependence measures, the numerical values of the corresponding measures change if we apply such a re-scaling. It is therefore desirable to come up with alternative measures of dependence which would not change under such transformations.

Copulas: brief reminder. A description of a joint probability distribution of quantities X and Y which does not change under arbitrary re-scaling is known as a *copula*; see, e.g., [5, 7]. A copula corresponding to the distribution $F(x, y)$ with marginal distributions $F_X(x)$ and $F_Y(y)$ can be defined as

$$C(u, v) = F\left(F_X^{-1}(u), F_Y^{-1}(v)\right), \tag{14}$$

where $F_X^{-1}(u)$ and $F_Y^{-1}(v)$ denote functions which are inverse to $F_X(x)$ and $F_Y(y)$.

One can easily check that this expression indeed does not change if we re-scale each of the quantities X and Y. Each copula is a probability distribution on the unit square $[0, 1] \times [0, 1]$ whose marginals are uniform distributions.

Copulas corresponding to the independent and varying independent cases. The case of two independent quantities corresponds to the copula $C(u, v) = u \cdot v$. For the varying independent case, from the formula (3), we conclude that

$$C(u, v) = \sum_{k=1}^{K} w_k \cdot a_k(u) \cdot b_k(v), \tag{15}$$

where $a_k(u) \stackrel{\text{def}}{=} A_k\left(F_X^{-1}(u)\right)$ and $b_k(v) \stackrel{\text{def}}{=} B_k\left(F_X^{-1}(v)\right)$.

How to detect varying independence based on the copula: idea. Since a copula is itself a probability distribution, we can use methods described in the previous sections to detect varying dependence based on a copula.

Algorithm for the idealized case. Let us first consider the ideal case, when for some grid values u_i and v_j ($1 \le i \le I$, $1 \le j \le J$), we know the exact values $C(u_i, v_j)$ of the corresponding copula. We want co check whether, for a given integer K, this copula corresponds to K-varying independence case, i.e., that this copula can be presented in the form (15).

For this checking, we consider vectors $\vec{C}_i \stackrel{\text{def}}{=} (C(u_i, b_1), \ldots, C(u_i, v_j))$ one by one.

We try $i' = 1, 2, \ldots$, and out of the first i' vectors we find the set S consisting of K linearly independent ones. We start with an empty set S.

- Once the set S is formed for some i', we check whether the next vector $\vec{C}_{i'+1}$ is linearly independent from S (see, e.g., [1]).
- If the vector $\vec{C}_{i'+1}$ is a linear combination of vectors from the set S, then we keep the set S intact and go to the next value i'.
- If $\vec{C}_{i'+1}$ is linearly independent from S, then we add this vector $\vec{C}_{i'+1}$ to the set S, and also go to the next value i'.

If at some point, we get a set S with $> K$ elements, we stop and conclude that a representation of type (15) is impossible, i.e., we have a true dependence. On the other hand, if after considering all I vectors $\vec{C}_1$, $\ldots$, $\vec{C}_I$, we have a set S with $\le K$ vectors, this means that a representation of type (15) is possible, i.e., we have a K-varying independence.

Algorithm for the case of an approximately known copula. Let us now assume that we know the copula values with a known accuracy σ. Then:

- We first compute the Frobenius norm of the matrix C with the components $C(u_i, v_j)$:
$$\|C\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{I} \sum_{j=1}^{J} (C(u_i, v_j))^2}. \tag{16}$$
- Then, we apply SVD to the matrix C to compute the first K singular values $\sigma_1, \ldots, \sigma_K$.
- After that, we compute
$$s_{\min} = \|C\|_F^2 - \sum_{k=1}^{K} \sigma_k^2. \tag{17}$$
- Finally, we compute $a = \dfrac{s_{\min}}{I \cdot J}$.

– If $a \le \sigma^2$, then the available information is consistent with the varying dependence.

– If $a > \sigma^2$, then we have true dependence.

Copula-based measures of dependence: definition. We arrive at the following definition.

Definition 2. *Let the values $C(u_i, v_j)$ of the copula be given, and let a positive integer K be given. By a* measure of deviation from K-varying independence*, we mean the value $c_K \stackrel{\text{def}}{=} \sqrt{\frac{s_{\min}}{I \cdot J}}$, where $s_{\min}$ is the smallest possible value of the quantity*

$$s \stackrel{\text{def}}{=} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(C(u_i, y_j) - \sum_{k=1}^{K} w_k \cdot a_k(u_i) \cdot b_k(v_j) \right)^2 \tag{18}$$

over all possible values of w_k, $a_k(u_i)$ and $b_k(v_j)$.

How to compute copula-based measures of dependence.

- First, we compute the Frobenius norm (16) of the matrix C with the components $C(u_i, v_j)$.
- Then, we apply SVD to the matrix C to compute the first K singular values $\sigma_1, \ldots, \sigma_K$.
- After that, we compute $s_{\min}$ by using the formula (17).
- Finally, we compute $c_K = \sqrt{\frac{s_{\min}}{I \cdot J}}$.

Acknowledgments

This work was supported by Grant No. P202/10/P360 from the Czech Science Foundation, by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and by a grant N62909-12-1-7039 from the Office of Naval Research.

This work was performed when M. Krmelova and M. Trnecka were visiting the University of Texas at El Paso. The authors are thankful to Hung T. Nguyen for valuable discussions.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2009.

[2] G. H. Golub and C. F. van Loan, *Matrix Computations*, John Hopkins Press, 2013.

[3] J. Hastad, "Tensor rank is NP-complete", *Journal of Algorithms*, 1990, Vol. 11, pp. 644–654.

[4] J. Hopcroft and R. Kannan, *Foundations of Data Science*, to appear.

[5] P. Jaworski, F. Durante, W. K. Härdle, and T. Ruchlik (eds.), *Copula Theory and Its Applications*, Springer Verlag, Berlin, Heidelberg, New York, 2010.

[6] T. G. Kolda and B. W. Bader, "Tensor Decompositions and Applications", *SIAM Reviews*, 2009, Vol. 51, No. 3, pp. 455–500.

[7] R. B. Nelsen, *An Introduction to Copulas*, Springer Verlag, Berlin, Heidelberg, New York, 1999.

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