

A note on symmetrically correlated random variables

Steffen Hoernig
UNL Lisboa

1998 (revision 2001)

Abstract

We show that there is a lower bound on the possible correlation between random variables that are identically distributed and pairwise symmetrically correlated. This lower bound is increasing in the number of random variables. While it is binding for example for normal distributions, we show that for multinomial Bernoulli distributions it is only obtained for certain values of the unconditional expectation, and that the true bound may even be arbitrarily close to zero.

1 Introduction

We show that there is a lower bound on the correlation between symmetrically correlated identically distributed random variables. As is intuitive, this lower bound is equal to -1 with two random variables, and converges to zero as the number of random variables approaches infinity (The upper bound on (positive) correlation is obviously $+1$, i.e. perfect correlation). This theoretical lower bound can be strictly lower than the smallest values of the correlation coefficient obtained for some specific distributions, where it is even possible that the lowest one must be arbitrarily close to zero.

2 A lower bound on correlation

Let there be n identically distributed and symmetrically correlated random variables. Expectations, covariances and variances, respectively, exist and are equal. The correlation coefficient between any two random variables is equal to ρ . This situation is very different from correlation in time series, which decreases over time. The variance-covariance matrix of these n random variables is

$$V_n = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix},$$

with σ^2 on the diagonal and $\rho\sigma^2$ everywhere off-diagonal. Without loss of generality let $\sigma^2 = 1$.

It is well-known that $D_n = \det V_n = (1 + (n - 1)\rho)(1 - \rho)^{n-1}$. Since V_n must be positive definite, it is necessary that

$$\rho_* = -\frac{1}{n-1} < \rho < 1. \tag{1}$$

This is already the basic result ... nothing deep.

Now the question is: Is this lower bound the strictest one possible, i.e. the minimum of feasible ρ , or are there tighter lower bounds? We will answer this question in the next two corollaries, and the answer is: Yes and no, it depends on the distribution. There are distributions where this bound holds strictly, and there are others where it doesn't.

Consider the multivariate Normal distribution: It is defined if and only if the expectations and the variance-covariance matrix of the random variables exist. Since the expectations are not affected by correlation, and the

variance-covariance matrix exists whenever $\rho \in (\rho_*, 1)$, we immediately get the following corollary:

Corollary 1 *There are joint distributions for which ρ_* is the smallest lower bound on ρ for all possible values of other parameters. Thus there is no strictly higher lower bound for all distributions.*

And now for the "yes and no" result:

Lemma 2 *For n identically distributed and symmetrically correlated multinomial Bernoulli random variables the minimum of feasible ρ is equal to ρ_* , and is taken on only at the unconditional expectations*

$$E[X] = \frac{i}{n}, \quad i = 1 \dots n - 1.$$

For all other values of $E[X]$ the minimum is strictly higher than ρ_ , and converges to zero as $E[X] \rightarrow 0$ or $E[X] \rightarrow 1$.*

Proof. See the next section. ■

From this follows:

Corollary 3 *There are joint distributions where depending on the expectation the lower limit on ρ is strictly higher than ρ_* , or even arbitrarily close to zero.*

3 The Multinomial Bernoulli case

3.1 Introduction

Let the random variables X_1, \dots, X_n have identical Bernoulli distributions with expected value $E[X_k] = v$, and be symmetrically correlated. Let their joint distribution be described by the probabilities x_0, \dots, x_n , where

$$x_i = P(X_1 = \dots = X_i = 0, X_{i+1} = \dots = X_n = 1).$$

Then it is easy to see that

$$1 = \sum_{i=0}^n \binom{n}{i} x_i \tag{2}$$

$$v = P(X_n = 1) = \sum_{i=0}^{n-1} \binom{n-1}{i} x_i, \tag{3}$$

and define

$$p = P(X_n X_{n-1} = 1) = \sum_{i=0}^{n-2} \binom{n-2}{i} x_i. \quad (4)$$

Consider the marginal distribution of (X_{n-1}, X_n) , with probabilities $(p, v-p, v-p, 1+p-2v)$. The variance of each is $V = v(1-v)$, and their covariance is

$$\begin{aligned} C &= (1-v)^2 p - 2(1-v)v(v-p) + v^2(1+p-2v) \\ &= p - v^2. \end{aligned}$$

The correlation coefficient is $\rho = (p - v^2)/v(1-v)$. Therefore we have obtained our first result:

Remark 4 *To find the minimum correlation coefficient it is necessary and sufficient to find the minimum values that p can take on.*

In particular, since ρ only depends on v and p , the only restrictions we must consider are (2), (3) and (4).

3.2 The Minimum Correlation

To find the minimum of p given the unconditional expectation v , we set up the following linear program:

$$\begin{aligned} p^* &= \min \sum_{i=0}^{n-2} \binom{n-2}{i} x_i \\ &s.t. \sum_{i=0}^n \binom{n}{i} x_i = 1 \\ &\quad \sum_{i=0}^{n-1} \binom{n-1}{i} x_i = v \\ &\quad x_0, \dots, x_n \geq 0 \end{aligned} \quad (5)$$

We will see some examples for $n = 3$, where the lower limit is $\rho^* = -\frac{1}{n-1} = -\frac{1}{2}$.

$$\begin{aligned} \min & x_0 + x_1 \\ & x_0 + 3x_1 + 3x_2 + x_3 = 1 \\ & x_0 + 2x_1 + x_2 = v \\ & x_0, x_1, x_2, x_3 \geq 0 \end{aligned}$$

$$\begin{aligned}
v = 1/3: & x_0 = 0, x_1 = 0, x_2 = \frac{1}{3}, x_3 = 0, p = 0, \rho = \frac{0 - (1/3)^2}{(1/3)(1 - 1/3)} = -\frac{1}{2} \\
v = 1/2: & x_0 = 0, x_1 = \frac{1}{6}, x_2 = \frac{1}{6}, x_3 = 0, p = \frac{1}{6}, \rho = \frac{1/6 - (1/2)^2}{(1/2)(1 - 1/2)} = -\frac{1}{3} \\
v = 2/3: & x_0 = 0, x_1 = \frac{1}{3}, x_2 = 0, x_3 = 0, p = \frac{1}{3}, \rho = \frac{1/3 - (2/3)^2}{(2/3)(1 - 2/3)} = -\frac{1}{2} \\
v = 99/100: & x_0 = \frac{97}{100}, x_1 = \frac{1}{100}, x_2 = 0, x_3 = 0, p = \frac{1}{100}, \\
\rho = & \frac{(98/100) - (99/100)^2}{(99/100)(1 - 99/100)} = -\frac{1}{99}.
\end{aligned}$$

Scientific Workplace minimization program: (change value of v and select Maple/Simplex/Minimize)

$$\begin{aligned}
& x_0 + x_1 \\
x_0 + 3x_1 + 3x_2 + x_3 &= 1 \\
x_0 + 2x_1 + x_2 &= v \\
x_0 &\geq 0 \\
x_1 &\geq 0 \\
x_2 &\geq 0 \\
x_3 &\geq 0
\end{aligned}$$

We see the following: It seems that the minimum $\rho^* = -1/(n-1)$ is achieved if and only if $v = i/n$ for $i = 1, \dots, n-1$. In the following we will first verify that the minimum is indeed achieved at these values of v , and then argue that these are the only ones.

Remark 5 *If $v = i/n$ for some $i = 1, \dots, n-1$ then the minimum is taken on at $x_{n-i} = 1/\binom{n}{n-i}$ and $x_j = 0$ for $j \neq n-i$.*

Proof. We first verify the constraints (2) and (3):

$$\begin{aligned}
\binom{n}{n-i} x_i &= \binom{n}{n-i} / \binom{n}{n-i} = 1 \\
\binom{n-1}{n-i} x_i &= \frac{(n-1)!}{(n-i)!(n-1-n+i)!} / \frac{n!}{(n-i)!(n-n+i)!} = \frac{i}{n}.
\end{aligned}$$

The value of the objective is

$$p = \binom{n-2}{n-i} x_{n-i} = \frac{(n-2)!}{(n-i)!(n-2-n+i)!} / \frac{n!}{(n-i)!(n-n+i)!} = \frac{i(i-1)}{n(n-1)},$$

giving rise to a correlation coefficient of

$$\rho = \frac{\frac{i(i-1)}{n(n-1)} - \left(\frac{i}{n}\right)^2}{\frac{i}{n} \left(1 - \frac{i}{n}\right)} = -\frac{1}{n-1}. \blacksquare$$

We will now show that for other values of v the minimum is not taken on. In order to do this we will analyse the dual of the above minimization program: Let s_1 and s_2 be the shadow variables of the constraints (2) and (3), respectively. Then the dual is

$$\begin{aligned}
& \max s_1 + vs_2 \\
& \binom{n}{0}s_1 + \binom{n-1}{0}s_2 \leq \binom{n-2}{0} \\
& \dots \binom{n}{i}s_1 + \binom{n-1}{i}s_2 \leq \binom{n-2}{i} \dots \\
& \binom{n}{n-3}s_1 + \binom{n-1}{n-3}s_2 \leq \binom{n-2}{n-3} \\
& \frac{n(n-1)}{2}s_1 + (n-1)s_2 \leq 1 \\
& ns_1 + s_2 \leq 0 \\
& s_1 \leq 0
\end{aligned}$$

This can be simplified to

$$\begin{aligned}
& \max s_1 + vs_2 \tag{6} \\
s_1 + \frac{n-i}{n}s_2 \leq \frac{(n-i)(n-1-i)}{n(n-1)}, \quad i = 0 \dots n
\end{aligned}$$

This dual has three very useful features: First, it has only two variables, which makes its solution intuitive. Second, the constraint set does not depend on v , while the objective does. Therefore varying v simply involves "sliding" a different objective along the upper right border of the constraint set. Third, the value of the dual's objective at the maximum is equal to the value of the primal's objective at the minimum, $\max s_1 + vs_2 = \min p$. Therefore we can choose which program to solve, and the choice is the dual.

Next step: We show that only the corners created by the intersections of two *neighboring* constraints matter. The point of intersection of any two constraints $i < j$ (corner) is

$$s_1 = -\frac{(n-j)(n-i)}{n(n-1)}, \quad s_2 = \frac{2n-1-i-j}{n-1}.$$

It can be shown that this corner fulfills restriction $k \neq i, j$ if $(k-j)(k-i) \geq 0$, i.e. either k is larger than j or smaller than i . On the other hand, if $i < k < j$, then condition k is violated. Therefore the corners of the feasible sets must belong to intersections of the conditions i and $i+1$, $i = 0, n-1$, with corner i given by

$$s_1 = -\frac{(n-i)(n-1-i)}{n(n-1)}, \quad s_2 = 2\frac{n-1-i}{n-1}.$$

We will now see where the objective $p = s_1 + vs_2$ touches the constraint set. It is easily seen that it touches corner i (which then is an optimal solution!) if and only if

$$\frac{n-i-1}{n} \leq v \leq \frac{n-i}{n}.$$

The values of the objective at corner i is

$$p = s_1 + vs_2 = \left(2v - \frac{n-i}{n}\right) \frac{n-1-i}{n-1},$$

which is the joint optimal value of the primal and the dual. If $i = 1, \dots, n-2$, at $v = \frac{n-i-1}{n}$ we obtain

$$\begin{aligned} p &= \frac{(n-2-i)(n-1-i)}{(n-1)n} \\ \rho &= \frac{\frac{(n-2-i)(n-1-i)}{(n-1)n} - \left(\frac{n-i-1}{n}\right)^2}{\left(\frac{n-i-1}{n}\right) - \left(\frac{n-i-1}{n}\right)^2} = -\frac{1}{n-1}, \end{aligned}$$

and at $v = \frac{n-i}{n}$,

$$\begin{aligned} p &= \frac{(n-i)(n-1-i)}{(n-1)n} \\ \rho &= \frac{\frac{(n-i)(n-1-i)}{(n-1)n} - \left(\frac{n-i}{n}\right)^2}{\left(\frac{n-i}{n}\right) - \left(\frac{n-i}{n}\right)^2} = -\frac{1}{n-1}, \end{aligned}$$

as expected.

The two border cases $i = 0$ and $i = n-1$ deserve special attention: If $i = n-1$ we obtain $p = 0$ at both $v = 0$ and $v = 1/n$, with $\rho = -1/(n-1)$ for the latter. At $v = 0$ the variance is zero, and therefore the correlation coefficient is not defined. Since $p = 0$ for all $v < 1/n$ we can take the limit, and arrive at:

$$\lim_{v \rightarrow 0} \frac{0 - v^2}{v - v^2} = \lim_{v \rightarrow 0} \left(-\frac{v}{1-v}\right) = 0.$$

Therefore, as v approaches 0, the correlation coefficient ρ approaches 0, and ρ is decreasing on $[0, 1/n]$. At $i = 0$, we have $\rho = -1/(n-1)$ at $v = (n-1)/n$, $p = 2v - 1$ for $v \geq (n-1)/n$ and for $v \in \left[\frac{n-1}{n}, 1\right]$

$$\rho = \frac{(2v-1) - v^2}{v - v^2} = \frac{v-1}{v},$$

where $\rho \rightarrow 0$ as $v \rightarrow 1$ and ρ is increasing on $[\frac{n-1}{n}, 1]$. Since ρ is zero when the random variables are uncorrelated this is the lowest upper limit on ρ , and we see that it is binding for $v \rightarrow 0$ and $v \rightarrow 1$.

We now examine the correlation coefficient ρ when $\frac{n-i-1}{n} < v < \frac{n-i}{n}$ for $i = 1, \dots, n-2$. Taking the derivative,

$$\begin{aligned} \frac{d}{dv}\rho &= \frac{d}{dv} \left(\frac{(2v - \frac{n-i}{n}) \frac{n-1-i}{n-1} - v^2}{v(1-v)} \right) \\ &= \frac{v^2 n(n-2i-1) - 2v(n-i)(n-1-i) + (n-i)(n-1-i)}{(n-1)v^2(1-v)^2 n}, \end{aligned}$$

we see that there are at most two critical values in the interval $\frac{n-i-1}{n} \leq v \leq \frac{n-i}{n}$ since the numerator is quadratic. Evaluating $d\rho/dv$ at the border values, we obtain $d\rho/dv > 0$ at the left border, and $d\rho/dv < 0$ at the right border. Therefore there is exactly one critical value in the interval, which must be a local maximum since $d\rho/dv$ is decreasing. Therefore the border values are local minima over $\frac{n-i-1}{n} \leq v \leq \frac{n-i}{n}$, and ρ is strictly higher for any v that is not a border value.

For curiosity: The local maximum for $\frac{n-i-1}{n} \leq v \leq \frac{n-i}{n}$ is obtained at

$$v = \frac{(n-i-1)(n-i) - \sqrt{i(1+i)(n-i-1)(n-i)}}{n(n-2i-1)},$$

with corresponding correlation coefficient

$$\rho = 2 \frac{i(n-i-1) - \sqrt{i(1+i)(n-i-1)(n-i)}}{n(n-1)} < 0.$$

Therefore:

Remark 6 $\rho = \rho^*$ if and only if $v = i/n$ for some $i = 1, \dots, n-1$. In fact, $\rho \rightarrow 0$ as $v \rightarrow 0$ or $v \rightarrow 1$, i.e. depending on the value of v negative correlation almost disappears.