On the Minimum Correlation between Symmetrically Distributed Random Variables

Steffen Hoernig*
Nova School of Business and Economics

June 9, 2018

Abstract

Using linear programming, we show that families of symmetrically distributed Bernoulli random variables have a maximal negative correlation that almost always is strictly above the general lower limit.

Keywords: Bernoulli random variables, Correlation coefficient
JEL: C10, C61

*Universidade Nova de Lisboa, Campus de Campolide, 1099-032 Lisboa, Portugal, shoernig@novasbe.pt.

1 Minimum Correlation

Consider $n \geq 2$ random variables $X_1, ..., X_n$ with a symmetric joint distribution function, i.e. permuting its arguments does not change its value. These random variables are symmetrically distributed, with identical marginal distributions and identical pairwise correlation. Denote the variance of each random variable as σ^2 and the pairwise correlation coefficient as r. The variances of the sum of all, and of the difference between a pair $i \neq j$, of these random variables are

$$V\left[\sum_{i=1}^{n} X_i\right] = n\left(1 + (n-1)\, r\right)\sigma^2, \quad V\left[X_i - X_j\right] = 2\left(1 - r\right)\sigma^2.$$

For these variances to be non-negative, it is necessary that the correlation coefficient lies in the interval

$$-\frac{1}{n-1} \leq r \leq 1. \tag{1}$$

These are also the exact same conditions that guarantee that the variance-covariance matrix of $X_1, ..., X_n$ is positive semi-definite. These guarantee that the variance of *any* linear combination of $X_1, ..., X_n$ is non-negative. The above sum and differences are special cases which correspond to the eigenvectors of the variance-covariance matrix. These facts are known, see e.g. [2], though not widely so.

It is also immediate to see that (1) provides the tightest *general* limits on the correlation coefficient: Joint normal distributions with exactly this type of variance-covariance matrix exist for all r in this interval.

The question we are posing in this paper is the following: Are there families of distributions for which the limit on negative correlation is strictly tighter? The answer is yes, and the example we explore is the Bernoulli distribution, i.e. $X_i \in \{0, 1\}$ with $P\left[X_i = 1\right] = p$. This distribution has one interesting aspect in common with the normal distribution, which is that two random variables are pairwise independent if and only if they are uncorrelated (this is not true in general: independence is a stronger notion). Still, the "lumpiness" of the Bernoulli distribution implies that in general the "most negative" symmetric correlation in families of these random variables is strictly higher than indicated by (1). We find the following:

Proposition 1 *For n jointly and symmetrically distributed Bernoulli random variables with expectation $p \in (0, 1)$, the minimum correlation coefficient, for $\frac{j}{n} \leq p \leq \frac{j+1}{n}$, $j = 0, ..., n-1$, is*

$$r_* = \frac{n}{n-1}\frac{\left(\frac{j+1}{n} - p\right)\left(p - \frac{j}{n}\right)}{p - p^2} - \frac{1}{n-1}.$$

In particular:

1. *The minimum correlation* $r_* = -\frac{1}{n-1}$ *is achieved if and only if* $p = \frac{i}{n}$, $i = 1, ..., n-1$, *while for all other* p *we have* $r_* > -\frac{1}{n-1}$.

2. *For* $p < \frac{1}{n}$ *we have* $r_* = -\frac{p}{1-p} \to_{p \to 0} 0$, *and for* $p > \frac{n-1}{n}$ *we have* $r_* = -\frac{1-p}{p} \to_{p \to 1} 0$.

3. *In each interval* $\frac{j}{n} \leq p \leq \frac{j+1}{n}$, $j = 1, ..., n-2$, *the maximal value of* r_* *is equal to* $r_* = -\frac{1}{n}$ *if* n *is odd and* $p = \frac{1}{2}$, *and otherwise*

$$r_* = -2\frac{\sqrt{j\,(j+1)\,(n-1-j)\,(n-j)} - j\,(n-1-j)}{n\,(n-1)},$$

at $p = \frac{\sqrt{j(j+1)(n-1-j)(n-j)} - j(1+j)}{n(n-1-2j)}$.

Proof. The proof is provided in the next section. We reformulate the search for the minimal correlation as the solution to a simple linear programme, with the probabilities of specific events as decision variables, and then solve its dual. ■

As an illustration, in Figure 1 we outline the minimum correlation coefficient r_* for the case $n = 3$. It is clearly visible that the general lower limit of $-\frac{1}{2}$ is only reached if either $p = \frac{1}{3}$ or $p = \frac{2}{3}$. For most other values of p, the limit r_* is significantly higher.

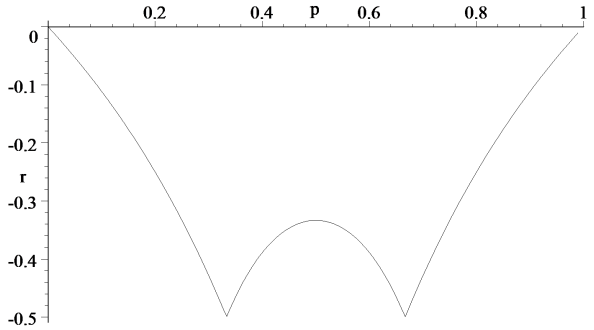

Figure 1: the minimum correlation coefficient r_* for $n = 3$.

We also provide an example of a distribution that achieves exactly r_*, for any given n: Fix some $j = 1, ..., n-1$ and let all outcomes have zero probability

unless $\sum_{i=1}^{n} X_i = n-j$. As there are $\binom{n}{n-j}$ of these outcomes, their individual probability is $x_j = 1/\binom{n}{n-j}$. Following the exposition below, we obtain

$$p = \binom{n-1}{n-j} x_j = \frac{(n-1)!}{(n-j)!\,(j-1)!} \Big/ \frac{n!}{j!\,(n-j)!} = \frac{j}{n},$$

i.e. $r_* = -\frac{1}{n-1}$.

2 The Proof

Let the random variables X_k, $k = 1, ..., n$, have identical Bernoulli distributions on $\{0, 1\}$ with expected value p and be symmetrically correlated. Their joint distribution is described by the probabilities $x_0, ..., x_n$, where

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

where the index i states the number of leading zeros in the ordered sequence $X_1, ..., X_n$ where $X_k = 1$ for all $k > i$. It is easy to see that

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

$$p = P\left(X_n = 1\right) = \sum_{i=0}^{n-1} \binom{n-1}{i} x_i. \tag{3}$$

Let

$$b = P\left(X_{n-1} = 1, X_n = 1\right) = \sum_{i=0}^{n-2} \binom{n-2}{i} x_i.$$

The marginal distribution of (X_{n-1}, X_n) then has probabilities $P(11) = b$, $P(01) = P(10) = p - b$ and $P(00) = 1 + b - 2p$, with covariance

$$\begin{aligned} Cov\left[X_{n-1}, X_n\right] &= (1-p)^2 b + 2(1-p)(0-p)(p-b) + (0-p)^2 (1+b-2p) \\ &= b - p^2. \end{aligned}$$

The correlation coefficient is therefore $r = \left(b - p^2\right)/p\,(1-p)$. As a result, in order to find the minimum correlation coefficient given the expectation p it is necessary and sufficient to find the minimum feasible value of b, subject to the conditions $x_0, ..., x_n \geq 0$, (2) and (3). We set up the following linear program:

$$b_* = \min_{x_0,...,x_n} \sum_{i=0}^{n-2} \binom{n-2}{i} x_i$$

$$s.t. \sum_{i=0}^{n} \binom{n}{i} x_i = 1, \quad \sum_{i=0}^{n-1} \binom{n-1}{i} x_i = p, \quad x_0, ..., x_n \geq 0$$

It is actually simpler to consider its dual. With s_1 and s_2 the shadow variables of the constraints (2) and (3), respectively, the dual problem is (see [1], ch. 4.2)

$$\max_{s_1,s_2} s_1 + ps_2 \ s.t. \ \binom{n}{i}s_1 + \binom{n-1}{i}s_2 \leq \binom{n-2}{i}, \quad i = 0, ..., n,$$

which can be restated in simpler form as

$$\max_{s_1,s_2} s_1 + ps_2 \quad s.t. \quad s_1 + \frac{n-i}{n}s_2 \leq \frac{(n-i)(n-1-i)}{n(n-1)}, \ i = 0 \ldots n.$$

This dual has three very useful features: First, it has only two variables, which makes its solution easy. Second, the constraint set does not depend on p. Therefore varying p simply involves sliding the objective along the upper right border of the constraint set. Third, since both programs have a finite solution, the value of the dual's objective at its maximum is equal to the value of the primal's objective at its minimum, $\max s_1 + ps_2 = b_*$.

It can be shown that the corners of the constraint set are given by the intersections of the neighboring constraints i and $i+1$, $i = 0, ..., n-1$, at coordinates $s_1^* = -\frac{(n-i)(n-1-i)}{n(n-1)}$, $s_2^* = 2\frac{n-1-i}{n-1}$. The objective $s_1 + ps_2$ touches the constraint set (and thus has an optimal solution) at corner $(i, i+1)$ if and only if $\frac{n-i-1}{n} \leq p \leq \frac{n-i}{n}$. To make this more intuitive, change the index to $j = n - i - 1$, for $j = 0, .., n-1$, so that this range becomes $\frac{j}{n} \leq p \leq \frac{j+1}{n}$. The value of the objective at the corresponding corner is then

$$b_* = s_1^* + ps_2^* = \frac{j}{n-1}\left(2p - \frac{j+1}{n}\right),$$

with correlation coefficient

$$r_* = \frac{b_* - p^2}{p - p^2} = \frac{n}{n-1}\frac{\left(\frac{j+1}{n} - p\right)\left(p - \frac{j}{n}\right)}{p - p^2} - \frac{1}{n-1}.$$

Clearly $r_* = -\frac{1}{n-1}$ at either $p = \frac{j}{n}$ or $p = \frac{j+1}{n}$, while $r_* > -\frac{1}{n-1}$ otherwise. Furthermore, for $j = 0$ this simplifies to $r_* = -\frac{p}{1-p}$, while for $j = n-1$ we have $r_* = -\frac{1-p}{p}$. Both converge to zero as p approaches 0 or 1, respectively.

In order to identify the locally maximal value of r_* on the interval $\frac{j}{n} \leq p \leq \frac{j+1}{n}$ for $j = 1, ..., n-2$, we take the derivative:

$$\frac{dr_*}{dp} = \frac{j(1+j)(1-2p) - n(n-1-2j)p^2}{n(n-1)p^2(1-p)^2}.$$

If n is odd then $\frac{dr_*}{dp} = 0$ at $p = \frac{1}{2}$, with $r_* = -\frac{1}{n}$. Evaluating $\frac{dr_*}{dp}$ at the border values, we obtain $\frac{dr_*}{dp} > 0$ at the left border and $\frac{dr_*}{dp} < 0$ at the right border. Since the numerator is quadratic in p there is exactly one critical value in the interval, which must be a local maximum. Solving $\frac{dr_*}{dp} = 0$, the local maximum is obtained at the values indicated above.

Acknowledgments

The core of this note was jotted down 20 years ago when I was still doing my Ph.D. at the EUI, Florence. Many thanks to the colleagues who over the years and many coffee breaks incited me to finally write it up. This work was funded by National Funds through FCT – Fundação para a Ciência e Tecnologia under the project Ref. UID/ECO/00124/2013 and by POR Lisboa under the project LISBOA-01-0145-FEDER-007722.

References

[1] Luenberger, David G. (1989). Linear and nonlinear programming, second edition, Reading, MA: Addison-Wesley.

[2] Vives, Xavier (2011). "Strategic supply function competition with private information", Econometrica, 79(6), 1919-1966.

Declarations of interest: none