Elastic Bounds of Pearson's r: Theoretical and Empirical Insights for Ordinal Data

Cees van der Eijk Scott Moser

Table of contents

Abstract

Pearson's correlation coefficient r, when applied to ordinal data, is subject to elasticity due to constraints from marginal distributions. This paper develops theoretical bounds for r, based on Boole–Fréchet–Hoeffding inequalities and rearrangement inequalities, and derives analytic expressions for their limits. We investigate when symmetry in the bounds holds or breaks, and explore practical implications for correlation-based methods like PCA and SEM. An accompanying R package allows users to compute bounds and perform hypothesis tests given fixed marginals.

1 Introduction and Motivation

- The widespread use of Pearson's r with ordinal data.
- The problem of elasticity: how max/min bounds of r vary with marginals.
- Contributions:
    1. Theory of optimal bounds
    2. Symmetry-breaking conditions
    3. Empirical consequences
    4. Supporting tools

2 Background and Definitions

- Ordinal data and assumptions of interval-scale correlation
- Existing alternatives: Spearman, polychoric, polyserial; their strengths and limitations
- Coupling and rearrangement in discrete settings

3 Theoretical Framework

3.1 Optimal Coupling and Correlation Bounds

We formalize the problem of identifying the joint distribution (coupling) of two ordinal variables with fixed marginals that maximizes or minimizes the Pearson correlation.

Proposition 1. Let X and Y be discrete ordinal random variables, each with finite ordered support:

$$X : x_1 < x_2 < \cdots < x_{K_X}, \quad Y : y_1 < y_2 < \cdots < y_{K_Y}$$

with marginal probability distributions $p_X(x_i) = P(X = x_i)$, $p_Y(y_j) = P(Y = y_j)$, respectively.

Then, the maximum and minimum values of the Pearson correlation coefficient r_{XY}, consistent with the fixed marginals, are obtained as follows:

- Maximum r_{XY}: achieved by pairing largest X-values with largest Y-values (comonotonic arrangement).

- Minimum r_{XY}: achieved by pairing largest X-values with smallest Y-values (anti-comonotonic arrangement).

Proof. The Pearson correlation coefficient is given by:

$$r_{XY} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}.$$

Given fixed marginals $p_X(x_i)$, $p_Y(y_j)$, the expectations $E[X], E[Y]$ and standard deviations σ_X, σ_Y are constant. Thus, to maximize or minimize r_{XY}, we only need to maximize or minimize the expectation $E[XY]$:

$$E[XY] = \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} x_i y_j P(X = x_i, Y = y_j).$$

This proof relies on a rearrangement inequality of @hard52-inequalities (Theorem 368):

For real sequences $a_1 \le a_2 \le \dots \le a_n$ and $b_1 \le b_2 \le \dots \le b_n$, and for any permutation π of indices $\{1, \dots, n\}$ we have the following:

$$a_1 b_1 + a_2 b_2 + \dots + a_n b_n \ge a_1 b_{\pi(1)} + a_2 b_{\pi(2)} + \dots + a_n b_{\pi(n)}$$

and

$$a_1 b_n + a_2 b_{n-1} + \dots + a_n b_1 \le a_1 b_{\pi(1)} + a_2 b_{\pi(2)} + \dots + a_n b_{\pi(n)}$$

Equality occurs only if the permutation π is monotonically increasing (for maximum) or monotonically decreasing (for minimum).

To explicitly connect this to our problem, enumerate observations explicitly from sorted marginals:

$$X_\downarrow : x_{[K_X]} \ge x_{[K_X-1]} \ge \dots \ge x_{[1]}, \quad Y_\downarrow : y_{[K_Y]} \ge y_{[K_Y-1]} \ge \dots \ge y_{[1]}.$$

(Sorted in descending order for maximization) * Expand discrete probability distributions into explicit observation-level sequences. If $K_X \ne K_Y$, extend the shorter sequence with zero-probability categories so both sequences have equal length, which does not affect marginal probabilities or correlation.

Hence, to maximize $E[XY]$: pair the largest ranks of X with the largest ranks of Y. Explicitly, start from the top (highest values) and move downward:

- Pair as many observations of the largest X-category as possible with the largest Y-category, then next-largest, etc., until all observations are paired.

- By the Hardy–Littlewood–Pólya Rearrangement Inequality, this explicitly constructed pairing yields the maximum possible sum of products, hence maximizing $E[XY]$.

Likewise, to minimize $E[XY]$ pair the largest X-categories with the smallest Y-categories, then second-largest X-categories with second-smallest Y-categories, and so forth (anti-comonotonic ordering). By the rearrangement inequality, this arrangement explicitly yields the minimal sum of products, hence minimizing $E[XY]$.

□

In words, given fixed marginals, we have shown:

- Maximum correlation $r_{\max}$ is achieved by comonotonic arrangement:

$$P_{\max}(X = x_{[i]}, Y = y_{[i]}) \quad \text{in descending order.}$$

- Minimum correlation $r_{\min}$ is achieved by anti-comonotonic arrangement:

$$P_{\min}(X = x_{[i]}, Y = y_{[n+1-i]}) \quad \text{pairing descending X with ascending Y.}$$