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

Cees van der Eijk Scott Moser

Table of contents

Abstract

The use of product-moment correlations to assess relationships between ordered-categorical variables is widespread and generally tolerated in spite of the non-metric character of such variables. Potential problems of this practice have been identified in the literature. This article focusses on one of these, namely that correlation coefficients are poorly comparable across pairs of ordered-categorical variables because their ranges are constrained in non-uniform ways. This not only affects the validity of correlation coefficients for descriptive purposes, but also as bases for subsequent analyses, as in, e.g., principal component analysis, factor analysis, and structural equation modelling. This article focusses on this problem of 'constrained' correlations. It explains the problem and explores conditions under which it is most severe. It provides illustrations based on both fabricated and actual empirical data, the latter from a mass-survey that ubiquitously contains many ordered-categorical variables that are commonly analysed with product-moment correlations. The article is accompanied by a software tool in R for analysing the attainable upper- and lower bounds of correlations in actual data.

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 $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \leq b_2 \leq \cdots \leq b_n$, and for any permutation π of indices $\{1, \ldots, n\}$ we have the following:

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

and

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

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:

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

(Sorted in descending order for maximization) * Expand discrete probability distributions into explicit observation-level sequences. If $K_X \neq K_Y$, 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:

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

With this explicit construction we can simply calculate r_{max} and r_{min} when the values of X and Y are ranks.

Proposition 2 (Corollary: Analytic Forms for $r_{\min}$ and $r_{\max}$). Let two ordinal variables X and Y have values as follows:

* X with K_X ordinal levels: $1, 2, \ldots, K_X$

* Y with K_Y ordinal levels: $1, 2, \ldots, K_Y$

with absolute frequencies:

- $n_X(i)$, number of observations at level i of X.

- $n_Y(j)$, number of observations at level j of Y.

Then the maximum Pearson correlation is:

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

Where

$$E[XY]_{\max} = \frac{1}{N} \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} x_i y_j M_{i,j}$$

where explicitly defined frequencies $M_{i,j}$ pair observations in a comonotonic manner, calculated explicitly via the greedy iterative approach as follows.

Define the matrix $M_{i,j}$, the frequency with which the level x_i pairs with the level y_j. Clearly, we must have:

$$\sum_j M_{i,j} = n_X(i), \quad \sum_i M_{i,j} = n_Y(j)$$

Algorithm for explicit calculation of $M_{i,j}$ (greedy method):

Initialize available frequencies explicitly: $n'_Y(j) = n_Y(j) \quad$ for each j

- For each level x_i, from smallest $i = 1$ to largest:

- For each level y_j, from smallest $j = 1$ to largest:

- Pair as many observations as possible:

$$M_{i,j} = \min \left(n_X(i) - \sum_{s=1}^{j-1} M_{i,s},\, n'_Y(j) \right)$$

Next, update explicitly:

$$n'_Y(j) \leftarrow n'_Y(j) - M_{i,j}$$

—

Similarly, the minimum Pearson correlation ($r_{\min}$) achieved by anti-comonotonic alignment (pairing largest ranks of X with smallest ranks of Y) is explicitly:

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

Proof. Let the total number of observations be N, thus:

$$\sum_{i=1}^{K_X} n_X(i) = N, \quad \sum_{j=1}^{K_Y} n_Y(j) = N.$$

Means and standard deviations clearly become:

Means:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{K_X} i \cdot n_X(i), \quad \bar{Y} = \frac{1}{N} \sum_{j=1}^{K_Y} j \cdot n_Y(j).$$

Standard deviations:

$$\sigma_X = \sqrt{\frac{1}{N} \sum_{i=1}^{K_X} (i - \bar{X})^2 n_X(i)}, \quad \sigma_Y = \sqrt{\frac{1}{N} \sum_{j=1}^{K_Y} (j - \bar{Y})^2 n_Y(j)}.$$

To achieve the maximum expected product $E[XY]_{\max}$, do the following: First *sort* ranks from largest to smallest for both X and Y; then *pair* ranks directly from highest to lowest available observations.

Minimum Expectation

Define explicitly a matrix $R_{i,j}$ giving frequencies of pairings $X = x_i$ with $Y = y_j$. For the minimal expectation, pair largest available X-values explicitly with smallest available Y-values first (greedy algorithm):

Initialize explicitly available frequencies:

$$n'_Y(j) = n_Y(j) \quad \text{for each } j$$

* For i from largest to smallest ($i = K_X$ down to 1):

* For j from smallest to largest ($j = 1$ up to K_Y):

* Pair explicitly as many observations as possible:

$$R_{i,j} = \min \left(n_X(i) - \sum_{s=1}^{j-1} R_{i,s},\, n'_Y(j) \right)$$

Update explicitly available frequencies:

$$n'_Y(j) \leftarrow n'_Y(j) - R_{i,j}$$

Then explicitly the minimal expectation is:

$$E[XY]_{\min} = \frac{1}{N} \sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} x_i y_j R_{i,j}$$

This explicit equation applies Whitt's rearrangement theorem (@whit76-bivariate), ensuring proper maximal pairing of ranks.

The formula for $E[XY]_{\min}$ is derived in a similar maner.

□

Discussion

We want to construct a joint distribution table M such that:

- The row sums match the marginal counts for X: $\sum_j M_{i,j} = n_X(i)$
- The column sums match the marginal counts for Y: $\sum_i M_{i,j} = n_Y(j)$

To do this, we go greedily, pairing the highest available X values with the smallest available Y values, one cell at a time.

To avoid exceeding the available marginal totals in Y, we must track how much of $n_Y(j)$ is still unassigned at each step — and that's what $n'_Y(j)$ represents.
After assigning $M_{i,j}$ units to the cell (i, j), we subtract that amount from the remaining pool of level y_j values, like this: $n'_Y(j) \leftarrow n'_Y(j) - M_{i,j}$This ensures we don't overfill any column of the matrix.

To maximize the expectation $E[XY]$, the rearrangement inequality states we must pair values of X and Y from smallest-to-smallest, second-smallest-to-second-smallest, and so forth (comonotonic alignment).

Thus, explicitly:

- Sort X: lowest-to-highest:

$$X_{\text{sorted}} = (\underbrace{x_1, \ldots, x_1}_{n_X(1)}, \underbrace{x_2, \ldots, x_2}_{n_X(2)}, \ldots, \underbrace{x_{K_X}, \ldots, x_{K_X}}_{n_X(K_X)})$$

Sorted Y-values (ascending order for max, descending for min):

$$Y_{\text{sorted(max)}} = (\underbrace{y_1, \dots, y_1}_{n_Y(1)}, \underbrace{y_2, \dots, y_2}_{n_Y(2)}, \dots, \underbrace{y_{K_Y}, \dots, y_{K_Y}}_{n_Y(K_Y)})$$

$$Y_{\text{sorted(min)}} = (\underbrace{y_{K_Y}, \dots, y_{K_Y}}_{n_Y(K_Y)}, \underbrace{y_{K_Y-1}, \dots, y_{K_Y-1}}_{n_Y(K_Y-1)}, \dots, \underbrace{y_1, \dots, y_1}_{n_Y(1)})$$

The maximum and minimum expectations are simply:

- Maximum expectation (pair element-by-element):

$$E[XY]_{\max} = \frac{1}{N} \sum_{k=1}^{N} X_{\text{sorted}}(k) \cdot Y_{\text{sorted(max)}}(k)$$

- Minimum expectation:

$$E[XY]_{\min} = \frac{1}{N} \sum_{k=1}^{N} X_{\text{sorted}}(k) \cdot Y_{\text{sorted(min)}}(k)$$

3.2 Fréchet–Hoeffding and Boole–Fréchet Inequalities

These classical bounds define feasible joint distributions given marginals and constrain the attainable values of r.

3.3 Symmetry and the Role of Ties

- Conditions under which $r_{\min} = -r_{\max}$
- How ties in marginal distributions affect the attainability of the lower bound

3.4 Special Cases and Boundary Conditions

- Equal vs. unequal numbers of categories
- Symmetric vs. asymmetric marginal distributions

3.5 Relation to Copula Theory (Optional)

Although our setting is discrete, the problem is conceptually linked to copula-based modeling of dependence with fixed marginals. The connection is discussed as a direction for future work.

4 Empirical Illustration and Practical Relevance

4.1 Effects of Marginal Shapes

- Shape-based elasticity of correlation bounds
- Influence of skewness, modality, and entropy

4.2 Symmetry Breaking in Correlation Bounds

- When and why $r_{\min} \neq -r_{\max}$
- Empirical indicators of bound asymmetry

When we talk about "asymmetry" in this context, we're referring to how different the distribution of Y is from the distribution of X. Since X is kept uniform, any deviation of Y from uniformity creates asymmetry between the distributions. The total variation distance (TV) quantifies this asymmetry:

- It measures how different Y is from being symmetric around its midpoint
- When TV = 0: Y is symmetric (like X)
- When TV is large: Y is highly asymmetric compared to X

4.3 Applications and Implications

- Impact on PCA, factor analysis, and SEM using ordinal data
- Practical risks: over-factoring, biased structural coefficients
- Recommendations for interpreting Pearson's r in applied contexts

5 Extensions and Open Questions

- Role of entropy and joint uncertainty in bounding behavior
- Probabilistic modeling over the $r \in [r_{\min}, r_{\max}]$ interval
- Possibility of adjusting or normalizing r

6 Conclusion

- Summary: Theoretical bounds, symmetry-breaking conditions, and implications for applied research

- An R package accompanies this paper, providing tools to compute max/min bounds and perform randomization-based hypothesis tests for fixed marginals. Details are available in the appendix and online.

- Future work: deeper integration with copula models, correction strategies, and generalizations to higher dimensions

7 References

8 Appendix A: Mathematical Proofs

- Formal derivations of optimal coupling and analytical bounds

9 Appendix B: Software Tools and Code Listings

- Documentation and examples for the R package
- Description of randomization testing features

10 Appendix C: Additional Tables and Simulations

- Supplementary empirical examples, simulation results