Marginal Extremes: Defining the Limits of Association in Cross-Tabulated Data

Cees van der Eijk Scott Moser

Optimal Coupling

Proposition

Let X and Y be two ordinal random variables taking values in

$$\{1, 2, \ldots, K_X\} \quad \text{and} \quad \{1, 2, \ldots, K_Y\},$$

respectively. Suppose the marginal distributions of X and Y are given (i.e., $\Pr(X = i)$ is fixed for each $i \in \{1, \ldots, K_X\}$, and $\Pr(Y = j)$ is fixed for each $j \in \{1, \ldots, K_Y\}$). Among all joint distributions of (X, Y) consistent with these marginals, the maximum value of Spearman's rank-correlation is attained via a comonotonic ordering – by arranging the highest ranks of X with the highest ranks of Y. The minimum value of Spearman's rank-correlation is attained via a countermonotonic ordering – by arranging the highest ranks of X with the lowest ranks of Y.

Formally, if we define

$$\text{rank}(X) = \begin{cases} 1 & \text{if } X = 1, \\ 2 & \text{if } X = 2, \\ \vdots & \\ K_X & \text{if } X = K_X, \end{cases} \qquad \text{rank}(Y) = \begin{cases} 1 & \text{if } Y = 1, \\ 2 & \text{if } Y = 2, \\ \vdots & \\ K_Y & \text{if } Y = K_Y, \end{cases}$$

then the joint distribution that sorts X in descending order and Y in descending order (matching largest with largest, next-largest with next-largest, etc.) maximizes

$$\rho_S(X,Y) \;=\; \mathrm{corr}\big(\mathrm{rank}(X), \mathrm{rank}(Y)\big).$$

Proof

1. Spearman's Correlation as Pearson's Correlation of Ranks
   By definition, Spearman's rank-correlation $\rho_S(X,Y)$ is the Pearson correlation between $\mathrm{rank}(X)$ and $\mathrm{rank}(Y)$. That is,

   $$\rho_S(X,Y) \;=\; \frac{\mathrm{Cov}(\mathrm{rank}(X),\, \mathrm{rank}(Y))}{\sqrt{\mathrm{Var}(\mathrm{rank}(X))\;\mathrm{Var}(\mathrm{rank}(Y))}}.$$

   Maximizing ρ_S is equivalent to maximizing the expected product $\mathbb{E}[\mathrm{rank}(X)\,\mathrm{rank}(Y)]$ subject to the fixed marginal distributions of $\mathrm{rank}(X)$ and $\mathrm{rank}(Y)$.

2. Rewriting the Expectation
   Let $r_1 < r_2 < \cdots < r_{K_X}$ be the distinct rank values for X and $s_1 < s_2 < \cdots < s_{K_Y}$ be the distinct rank values for Y. (Here $r_i = i$ and $s_j = j$ in typical usage.) Any joint distribution $\Pr(X = i, Y = j)$ that respects $\Pr(X = i)$ and $\Pr(Y = j)$ must allocate probability mass in a 2D contingency table, but always summing to $p_X(i)$ in row i and $p_Y(j)$ in column j. The quantity to be maximized is

   $$\sum_{i=1}^{K_X} \sum_{j=1}^{K_Y} r_i\, s_j \;\Pr(X = i, Y = j).$$

3. Application of the Rearrangement Inequality
   The rearrangement inequality (Hardy–Littlewood–Pólya, Theorem 368) or the discrete analog by Whitt (1978) tells us that for two finite sequences $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_n\}$ (here, the sequences are effectively the rank values weighted by the probability masses), the sum of products $\sum a_i\, b_{\sigma(i)}$ is maximized precisely when both sequences are sorted in the same order (both ascending or both descending). In the probability setting, "sorting from largest to smallest" means that the highest ranks of X should be paired with the highest ranks of Y. Concretely, if X is in descending rank order $(K_X, K_X - 1, \ldots)$ and Y is also in descending rank order $(K_Y, K_Y - 1, \ldots)$, then the product of ranks $\mathrm{rank}(X)\,\mathrm{rank}(Y)$ is as large as possible in expectation.

4. Different Number of Levels
   If $K_X \neq K_Y$, the principle is the same. We let $\text{rank}(X) \in \{1, \ldots, K_X\}$ and $\text{rank}(Y) \in \{1, \ldots, K_Y\}$. The rearrangement inequality still applies: list all "mass points" of X in descending order of rank, and list all "mass points" of Y in descending order of rank. Pair them index-by-index so that the largest rank in X is matched with the largest rank in Y. This arrangement yields the maximal expected product of ranks.

5. Concrete Example (Different Levels)

   - Suppose $K_X = 3$ and $K_Y = 4$. Then $\text{rank}(X)$ takes values $\{1, 2, 3\}$ and $\text{rank}(Y)$ takes $\{1, 2, 3, 4\}$.
   - Let $\Pr(X = 3) = 0.2$, $\Pr(X = 2) = 0.5$, $\Pr(X = 1) = 0.3$; and $\Pr(Y = 4) = 0.1$, $\Pr(Y = 3) = 0.4$, $\Pr(Y = 2) = 0.3$, $\Pr(Y = 1) = 0.2$.
   - To maximize $\mathbb{E}[\text{rank}(X)\,\text{rank}(Y)]$, we sort X from 3 down to 1 and Y from 4 down to 1. We then fill the contingency table so that the 0.2 probability mass of $X = 3$ is paired as much as possible with the 0.1 mass of $Y = 4$, the 0.4 mass of $Y = 3$, and so on, always aligning the largest rank masses together.
   - This yields the comonotonic distribution that achieves the largest $\text{rank}(X) \cdot \text{rank}(Y)$ on average, and thus the maximum Spearman correlation.

6. Conclusion
   By the rearrangement argument, the maximal Spearman rank-correlation is obtained via the comonotonic (descending-with-descending) joint distribution. Ties or repeated categories do not affect this principle, other than allowing multiple solutions that achieve the same maximum. Thus, the proposition is proved.

Extremal Values

We seek to find the maximum Pearson correlation coefficient, $r_{\max}$, between two discrete variables X and Y that take values in $\{0, 1, 2, \ldots, K-1\}$, given their fixed marginal distributions.

Step 1: Define the Problem and Notation

We are given:

- $n_X = (n_{X=0}, n_{X=1}, ..., n_{X=K-1})$, where $n_{X=i}$ is the number of times $X = i$ appears.
- $n_Y = (n_{Y=0}, n_{Y=1}, ..., n_{Y=K-1})$, where $n_{Y=j}$ is the number of times $Y = j$ appears.
- The total number of observations:

$$N = \sum_{i=0}^{K-1} n_{X=i} = \sum_{j=0}^{K-1} n_{Y=j}$$

Step 2: Compute the Means and Standard Deviations

The mean values of X and Y are:

$$\bar{X} = \frac{1}{N} \sum_{i=0}^{K-1} i \cdot n_{X=i}, \quad \bar{Y} = \frac{1}{N} \sum_{j=0}^{K-1} j \cdot n_{Y=j}$$

The variances are:

$$\sigma_X^2 = \frac{1}{N} \sum_{i=0}^{K-1} (i - \bar{X})^2 \cdot n_{X=i}, \quad \sigma_Y^2 = \frac{1}{N} \sum_{j=0}^{K-1} (j - \bar{Y})^2 \cdot n_{Y=j}$$

Thus, the standard deviations are:

$$\sigma_X = \sqrt{\frac{1}{N} \sum_{i=0}^{K-1} (i - \bar{X})^2 \cdot n_{X=i}}, \quad \sigma_Y = \sqrt{\frac{1}{N} \sum_{j=0}^{K-1} (j - \bar{Y})^2 \cdot n_{Y=j}}$$

Step 3: Construct the Joint Distribution for Maximum Correlation

To maximize r, we must maximize:

$$\text{Cov}(X, Y) = E[XY] - \bar{X}\bar{Y}$$

To do this, we construct a Sorted Array:

1. Sort the values of X in descending order according to their frequencies.
2. Independently, sort the values of Y in descending order according to their frequencies.
3. Assign pairings (X_i, Y_i) in order, ensuring that the highest values of X are paired with the highest values of Y, while respecting the marginal totals.

Let $m_{i,j}$ denote the number of times the pair (i, j) appears. The optimal strategy follows:

$$m_{i,j} = \min(n_{X=i}, n_{Y=j})$$

This ensures that the highest available values of X and Y are paired together as much as possible.

Step 4: Compute $E[XY]$

Given the optimal pair assignments:

$$E[XY]_{\max} = \frac{1}{N} \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} i \cdot j \cdot m_{i,j}$$

Substituting $m_{i,j} = \min(n_{X=i}, n_{Y=j})$, we obtain:

$$E[XY]_{\max} = \frac{1}{N} \sum_{i=0}^{K-1} i \sum_{j=0}^{K-1} j \cdot \min(n_{X=i}, n_{Y=j})$$
