Correspondence Analysis (unsupervised learning)
(Discrete version of the canonical correlation analysis)

Marianna Bolla, Prof. DSc.

We have two – usually not independent – discrete (categorical) r.v.'s X and Y taking on n and m different values, respectively (e.g., hair-color and eye-color). The joint distribution R is estimated from the joint frequencies based on N observations:

$$r_{ij} = \frac{f_{ij}}{N} \qquad (i = 1, \dots, n;\ j = 1, \dots, m).$$

Let $\mathbf{R}$ denote the $n \times m$ matrix of r_{ij}'s. The marginal distributions P and Q are:

$$p_i = r_{i.} \quad (i = 1, \dots, n) \quad \text{and} \quad q_j = r_{.j} \quad (j = 1, \dots, m).$$

Let $\mathbf{P}$ and $\mathbf{Q}$ denote the $n \times n$ and $m \times m$ diagonal matrices containing the marginal probabilities in their main diagonals, respectively.

- **Problem:** to approximate the table $\mathbf{R}$ with a lower rank table. For this purpose we are looking for *correspondence factor pairs* α_l, β_l taking on values and having unit variance with respect to the marginal distributions, further being maximally correlated on certain constraints. $\mathbb{E}_R \alpha\beta$ is maximum (1) for the trivial (constantly 1) variable pair α_1, β_1. Then for $k = 2, \dots, \min\{n, m\}$ we are looking for the maximum of $\mathbb{E}_R \alpha\beta$ on the constraints that

$$\mathbb{E}_P \alpha = \mathbb{E}_Q \beta = 0, \quad \mathrm{Var}_P \alpha = \mathrm{Var}_Q \beta = 1, \quad \mathrm{Cov}_P \alpha\alpha_i = \mathrm{Cov}_Q \beta\beta_i = 0 \quad (i = 1, \dots, k-1).$$

- **Solution**: by the SVD of the $n \times m$ matrix $\mathbf{B} = \mathbf{P}^{-1/2}\mathbf{R}\mathbf{Q}^{-1/2} = \sum_{l=1}^r s_l \mathbf{v}_l \mathbf{u}_l^T$, where r is the rank of $\mathbf{R}$. The singular values $1 = s_1 \ge s_2 \ge \dots \ge s_r > 0$ are the correlations of the correspondence factor pairs, while the values taken on by the k-th pair are coordinates of the *correspondence vectors* $\mathbf{P}^{-1/2}\mathbf{v}_k$ and $\mathbf{Q}^{-1/2}\mathbf{u}_k$, respectively.

- **Hypothesis testing**: to test the independence of X and Y, the test statistic

$$t = N \sum_{i=1}^n \sum_{j=1}^m \frac{(r_{ij} - p_i q_j)^2}{p_i q_j} = N \left[\sum_{i=1}^n \sum_{j=1}^m \frac{r_{ij}^2}{p_i q_j} - 1 \right] = N \|\mathbf{B} - \mathbf{B}_1\|_F^2 = N \sum_{l=2}^r s_l^2$$

  is used, that under independence follows $\chi^2((n-1)(m-1))$-distribution. To test the hypothesis that a k-rank approximation of the contingency table suffices, the statistic

$$N \|\mathbf{B} - \mathbf{B}_k\|_F^2 = N \sum_{l=k+1}^r s_l^2, \qquad k = 1, \dots, r-1$$

  should take on „small values" (its distribution is yet unknown), where $\mathbf{B}_k = \sum_{l=1}^k s_l \mathbf{v}_l \mathbf{u}_l^T$ is the best k-rank approximation of $\mathbf{B}$ in Frobenius norm (in spectral norm, too).

- **Definition**: The above t is called *total correspondence* of the table, while $N \sum_{l=2}^k s_l^2 / t$ is called *correspondence explained by the first k correspondence factor pairs.*

- **Spacial representation** of the row and column categories in k dimension: by the points

$$(\alpha_1(i), \alpha_2(i), \dots, \alpha_k(i)) \quad \text{and} \quad (\beta_1(j), \beta_2(j), \dots, \beta_k(j))$$

  $(i = 1, \dots, n;\ j = 1, \dots, m)$. Store them for further analysis.

- The optimal k is obtained by inspecting the singular values. Given $k(\leq \text{rank}\mathbf{B})$, the best rank k approximation of $\mathbf{R}$ is the following $\mathbf{R}^{(k)}$:

$$r_{ij}^{(k)} = p_i q_j \left(1 + \sum_{l=2}^{k} s_l \alpha_l(i) \beta_l(j) \right)$$

  Explanation: consider the SVD

$$\mathbf{B} = \sum_{l=1}^{r} s_l \mathbf{u}_l \mathbf{v}_l^T,$$

  i.e., for the entries

$$b_{ij} = \sum_{l=1}^{r} s_l \mathbf{u}_l(i) \mathbf{v}_l(j).$$

  Therefore,

$$\frac{r_{ij}}{\sqrt{p_i}\sqrt{q_j}} = \sum_{l=1}^{r} s_l \sqrt{p_i} \alpha_l(i) \sqrt{q_j} \beta_l(j),$$

  consequently,

$$r_{ij} = p_i q_j \sum_{l=1}^{r} s_l \alpha_l(i) \beta_l(j) = p_i q_j \left(1 + \sum_{l=2}^{r} s_l \alpha_l(i) \beta_l(j) \right),$$

  where the coordinates $\alpha_l(i)$, $\beta_l(j)$ are the so-called *canonical scores* of the lth factor.