Applying Correspondence Analysis (CA) in an Economic Dataset

Benghenima hafsa - Ghandouz Amina - G2 IASD

1 Introduction:

Correspondence Analysis (CA) is a multivariate statistical technique used to analyze and visualize relationships within categorical data. In the field of economics, where datasets often contain qualitative variables such as sectors, regions, or types of expenditures, CA provides a powerful tool to uncover patterns, associations, and underlying structures that might not be immediately visible through simple tabular analysis.

Therefor, we applied Correspondence Analysis to an economic dataset from WITS in order to interpret the associations between the different categorical variables. By projecting the data into a lower-dimensional space, we are able to identify which categories are closely related and explore potential economic insights. The results are presented through graphical representations and interpreted in the context of the economic dimensions they reveal.

2 About the dataset:

The dataset we used is from the World Integrated Trade Solution (WITS) platform and focuses on economic and trade indicators across multiple countries. It contains 18,836 rows and 7 columns (after removing duplicated and null samples), covering trade data between reporting countries and their partners over the period 1995 to 2021 except the years (1996, 1997, 1999, 2001, 2002, 2004)

The main columns include:

- **Reporter**: The country reporting the data.
- **Year**: The year of the observation.
- **Partner**: The trade partner country.
- **Product categories**: The category of traded goods or services.
- **Indicator Type and Indicator**: The type and name of the economic metric (e.g., exports, imports, trade indexes).
- **Indicator Value**: The corresponding numeric value for the indicator.

The primary objective of this study is to apply Correspondence Analysis (CA) to explore the underlying structure and relationships between categorical variables such as reporter, partner, product category and indicator type

3 Correspondence Analysis

Correspondence Analysis (CA) is a multivariate statistical technique used to explore relationships between categorical variables in a contingency table. It transforms qualitative data into a visual map, allowing us to identify patterns, similarities, and associations between rows and columns. To apply CA, we first create a contingency table from two categorical variables, then decompose it using singular value decomposition (SVD) to extract components (dimensions). The results include coordinates for rows and columns, explained inertia (variance captured), and biplots that visually represent the associations.

4 Correspondence Analysis Results:

In this study, we applied CA to our dataset using three different pairs of categorical variables. The goal was to explore the associations between these variables and identify any underlying patterns or structures.

1. **CA between Product Categories and Indicator Type:**

    - Contingency Table Shape: (22, 5).
    - Explained Inertia:
        (a) Component 1: 62.7%.
        (b) Component 2: 37.3%.
    - This CA reveals how product categories are related to different indicator types. The higher inertia on the first component suggests that a significant portion of the variation is captured by the first dimension. The second component provides additional insight, though less significant, into the structure.

2. **CA between Partner and Product Categories:**

    - Contingency Table Shape: (157, 22).
    - Explained Inertia:
        (a) Component 1: 62.5%.
        (b) Component 2: 37.5%.
    - Here, we see the relationship between trade partners and product categories. Similar to the first CA, the first component captures most of the variation, while the second component offers supplementary insights into the trade patterns. This can help identify clusters of trade partners that focus on similar product categories.

3. **CA between Reporter and Partner:**

    - Contingency Table Shape: (204, 157).
    - Explained Inertia:
        (a) Component 1: 52.9%.
        (b) Component 2: 47.1%.
    - In the third application, the relationship between reporters (countries or regions) and their trading partners is analyzed. The inertia distribution is almost evenly split between the two components, indicating that both dimensions contribute almost equally to explaining the data structure. This suggests a complex relationship between reporters and partners, where both axes play an essential role in defining the trade landscape.

4.1 Interpretation of Results:

From these three analyses, we observe that in all cases, the first component captures the majority of the variation in the data, though the importance of the second component is still substantial. The results provide insights into how product categories, trade partners, and reporters (countries/regions) are related to one another.