Correspondence Analysis

ST 558: Multivariate Analytics

Module 9

Lecture 2

Correspondence Analysis: Overview

Correspondence analysis is another technique primarily used to graphically represent high-dimensional data; specifically, it is used to display *contingency table* data, most commonly in two dimensions.

The resulting representation gives an indication of

- Which rows are similar to each other
- Which columns are similar to each other
- Rows and columns that have higher than expected association with each other.

Correspondence Analysis: Background

Recall that a *contingency table* is a way of representing the relationship between two (or more) different *categorical* variables:

Variable 1	Variable 2				Row Totals
	Level 1	Level 2	⋯	Level c	
Level 1	n_{11}	n_{12}	⋯	n_{1c}	$n_{1\cdot}$
Level 2	n_{21}	n_{22}	⋯	n_{2c}	$n_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	
Level r	n_{r1}	n_{r2}	⋯	n_{rc}	$n_{r\cdot}$
Column Totals	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot c}$	n

Correspondence Analysis: Background

We will let $\mathbf{P}$ be the matrix of *relative frequencies*.

$$\mathbf{P} = \begin{array}{|c|c|c|c|} \hline n_{11}/n & n_{12}/n & \cdots & n_{1c}/n \\ \hline n_{21}/n & n_{22}/n & \cdots & n_{2c}/n \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline n_{r1}/n & n_{r2}/n & \cdots & n_{rc}/n \\ \hline \end{array}$$

Define

- $\mathbf{r} = \frac{1}{n}[n_{.1}, n_{.2}, \ldots, n_{.r}]^T$: the vector of *row frequencies*
- $\mathbf{c} = \frac{1}{n}[n_{1.}, n_{2.}, \ldots, n_{c.}]^T$: the vector of *column frequencies*

Correspondence Analysis: Background

$$\mathbf{r} = \frac{1}{n}\begin{bmatrix} n_{1\cdot} \\ n_{2\cdot} \\ \vdots \\ n_{r\cdot} \end{bmatrix}$$

Variable 1	Variable 2				Row Totals
	Level 1	Level 2	⋯	Level c	
Level 1	n_{11}	n_{12}	⋯	n_{1c}	$n_{1\cdot}$
Level 2	n_{21}	n_{22}	⋯	n_{2c}	$n_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	
Level r	n_{r1}	n_{r2}	⋯	n_{rc}	$n_{r\cdot}$
Column Totals	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot c}$	n

$$\mathbf{c} = \frac{1}{n}\begin{bmatrix} n_{\cdot 1} \\ n_{\cdot 2} \\ \vdots \\ n_{\cdot c} \end{bmatrix}$$

Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student

- What their favorite ice cream flavor is:
  - Chocolate, Vanilla, or Strawberry
- What their favorite sport is:
  - Baseball, Basketball, Football, Soccer, or Tennis

We obtain the following data:

	Baseball	Basketball	Football	Soccer	Tennis
Chocolate	11	20	17	11	11
Vanilla	6	17	19	7	17
Strawberry	18	8	8	17	13

Correspondence Analysis: Steps

Correspondence analysis proceeds via the following steps:

1. Calculate the *expected frequency matrix* (that is, the expected table if the rows and columns were *independent*)
2. Calculate the matrix of *deviations from independence*
3. Calculate the *generalized singular value decomposition* (generalized SVD) of the deviations from independence
4. Use the first several (typically 2) dimensions of the generalized SVD, **appropriately scaled**, as the coordinates for plotting the rows (columns, respectively) of the contingency table

Correspondence Analysis: Steps

1. Calculate the *expected frequency matrix* (that is, the expected table if the rows and columns were *independent*)
   - The *expected frequency matrix* if the row variable and column variable are independent of each other is given by

$$\widehat{\mathbf{P}} = \mathbf{r}\mathbf{c}^T$$

2. Calculate the matrix of *deviations from independence*

$$\mathbf{P} - \widehat{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}^T$$

Correspondence Analysis: Steps

1. Calculate the *expected frequency matrix* (that is, the expected table if the rows and columns were *independent*)
   - The *expected frequenc* ...d column variable are in... given by

$$\widehat{\mathbf{P}} = \mathbf{r}\mathbf{c}^T$$

2. Calculate the matrix of *deviations from independence*

$$\mathbf{P} - \widehat{\mathbf{P}} = \mathbf{P} - \mathbf{r}\mathbf{c}^T$$

This difference matrix tells us how *extreme* the individual cells in the contingency table are: elements with larger values in this deviations matrix are more unusual if the variables are independent.

Correspondence Analysis: Steps

3. Calculate the *generalized singular value decomposition* (generalized SVD) of the deviations from independence
   - The *generalized SVD* is a way of expressing the matrix $\mathbf{P} - \mathbf{r}\mathbf{c}^T$ as a matrix product:

$$\mathbf{P} - \mathbf{r}\mathbf{c}^T = \mathbf{A}\boldsymbol{\Lambda}\mathbf{B}^T$$

   where $\boldsymbol{\Lambda}$ is a diagonal matrix, and the columns of $\mathbf{A}$ and $\mathbf{B}$ are orthogonal to each other, and scaled so that the *norm* of each column is proportional to the sum of the corresponding row/column:

$$\mathbf{A}^T\mathbf{D}_r\mathbf{A} = \mathbf{B}^T\mathbf{D}_c\mathbf{B} = \mathbf{I}$$

Correspondence Analysis: Steps

3. Calculate the *generalized singular value decomposition* (generalized SVD) of the

$$\mathbf{D}_r = \text{diag}(\mathbf{r}) = \begin{bmatrix} n_1./n & 0 & \cdots & 0 \\ 0 & n_2./n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_r./n \end{bmatrix}$$

$$\mathbf{D}_c = \text{diag}(\mathbf{c}) = \begin{bmatrix} n_{.1}/n & 0 & \cdots & 0 \\ 0 & n_{.2}/n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{.c}/n \end{bmatrix}$$

and $\mathbf{B}$ are orthogonal to each other, and scaled so that the *norm* of each column is proportional to the sum of the corresponding row/column:

$$\mathbf{A}^T \mathbf{D}_r \mathbf{A} = \mathbf{B}^T \mathbf{D}_c \mathbf{B} = \mathbf{I}$$

Correspondence Analysis: Steps

3. Calcula
   decomp
   deviati
   - The *ge*

     $\mathbf{P} - \mathbf{r}\mathbf{c}^T$ as a matrix product:

$$\mathbf{P} - \mathbf{r}\mathbf{c}^T = \mathbf{A}\mathbf{\Lambda}\mathbf{B}^T$$

     where $\mathbf{\Lambda}$ is a diagonal matrix, and the columns of $\mathbf{A}$ and $\mathbf{B}$ are orthogonal to each other, and scaled so that the *norm* of each column is proportional to the sum of the corresponding row/column:

$$\mathbf{A}^T\mathbf{D}_r\mathbf{A} = \mathbf{B}^T\mathbf{D}_c\mathbf{B} = \mathbf{I}$$

The generalized SVD tells us the predominant patterns in the rows and columns of a matrix.

Like the eigendecomposition/PCA, the first few components (columns of and $\mathbf{A}$, $\mathbf{B}$, with corresponding diagonal elements of $\mathbf{\Lambda}$) of this matrix product can give us a good approximation to the matrix.

Correspondence Analysis: Steps

4. Use the first several (typically 2) dimensions of the generalized SVD, appropriately scaled, as the coordinates for plotting the rows (columns, respectively)

$$\mathbf{X} = \mathbf{D}_r^{-1}\mathbf{A}\boldsymbol{\Lambda}$$

$$\mathbf{Y} = \mathbf{D}_c^{-1}\mathbf{B}\boldsymbol{\Lambda}$$

Correspondence Analysis: Steps

4. Use the first several (typically 2) dimensions of the generalized SVD, appropriately scaled, as the coordinates for plotting the rows (columns, respectively)

$$\mathbf{X} = \mathbf{D}_r^{-1}\mathbf{A}\boldsymbol{\Lambda}$$

$$\mathbf{Y} = \mathbf{D}_c^{-1}\mathbf{B}\boldsymbol{\Lambda}$$

Columns of **X** give coordinates for plotting the *row* deviations.

Columns of **Y** give coordinates for plotting the *column* deviations.

Correspondence Analysis: Steps

4. Use the first several (typically 2) dimensions of the generalized SVD, appropriately scaled, as the coordinates for plotting the rows (columns, respectively)

$$\mathbf{X} = \mathbf{D}_r^{-1}\mathbf{A}\boldsymbol{\Lambda}$$

$$\mathbf{Y} = \mathbf{D}_c^{-1}\mathbf{B}\boldsymbol{\Lambda}$$

Typically we plot the first two columns of $\mathbf{X}$, with each point representing the corresponding row of $\mathbf{P} - \mathbf{r}\mathbf{c}^T$

Typically we plot the first two columns of $\mathbf{Y}$, with each point representing the corresponding column of $\mathbf{P} - \mathbf{r}\mathbf{c}^T$

Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student

- What their favorite ice cream flavor is:
  - Chocolate, Vanilla, or Strawberry
- What their favorite sport is:
  - Baseball, Basketball, Football, Soccer, or Tennis

	Baseball	Basketball	Football	Soccer	Tennis	Row Totals
Chocolate	11	20	17	11	11	70
Vanilla	6	17	19	7	17	66
Strawberry	18	8	8	17	13	64
Column Totals	35	45	44	35	41	200

Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student

- What their favorite ice cream flavor is
- What their favorite sport is

The relative frequency matrix is therefore:

	Baseball	Basketball	Football	Soccer	Tennis	Row Totals
Chocolate	0.055	0.100	0.085	0.055	0.055	0.350
Vanilla	0.030	0.085	0.095	0.035	0.085	0.330
Strawberry	0.090	0.040	0.040	0.085	0.065	0.320
Column Totals	0.175	0.225	0.22	0.175	0.205	1.000

Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student

- What their favorite ice cream flavor is
- What their favorite sport is

The *expected* relative frequency matrix is therefore:

	Baseball	Basketball	Football	Soccer	Tennis	Row Totals
Chocolate	0.061	0.079	0.077	0.061	0.072	0.350
Vanilla	0.058	0.074	0.073	0.058	0.068	0.330
Strawberry	0.056	0.072	0.070	0.056	0.066	0.320
Column Totals	0.175	0.225	0.22	0.175	0.205	1.000

Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student

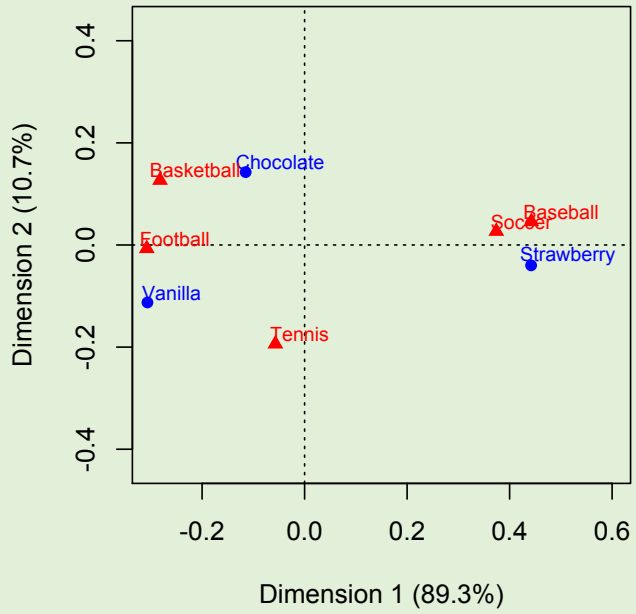
- What their favorite ice cream flavor is
- What their favorite sport is

The matrix of *deviations from independence* is therefore:

	Baseball	Basketball	Football	Soccer	Tennis	Row Totals
Chocolate	-0.006	0.021	0.008	-0.006	-0.017	0
Vanilla	-0.028	0.011	0.022	-0.023	0.017	0
Strawberry	0.034	-0.032	-0.030	0.029	-0.001	0
Column Totals	0	0	0	0	0	0

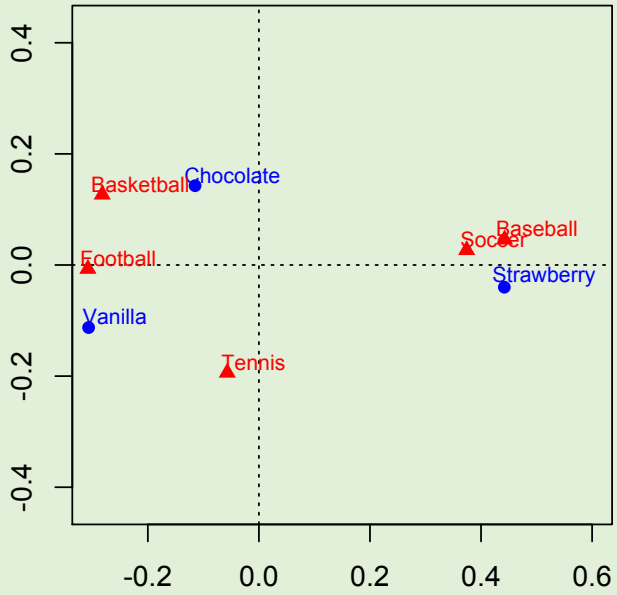
Correspondence Analysis: Example

Example: Suppose that we sample $n = 200$ OSU students and ask each sampled student what their favorite ice cream flavor is and what their favorite sport is. The resulting correspondence analysis plot is:



Correspondence Analysis: Example

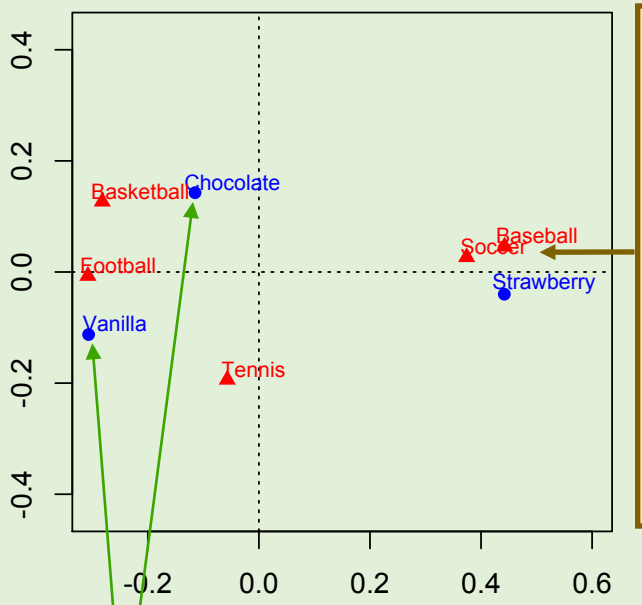
Example:



	Baseball	Basketball	Football	Soccer	Tennis
Chocolate	11	20	17	11	11
Vanilla	6	17	19	7	17
Strawberry	18	8	8	17	13

Correspondence Analysis: Example

Example:

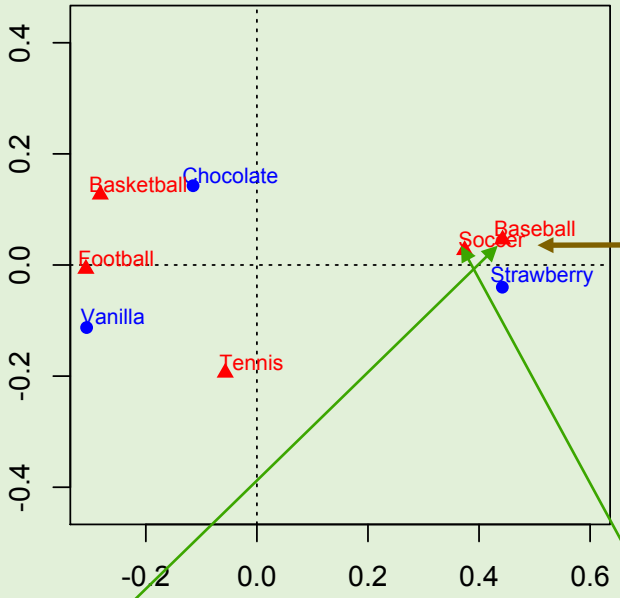


The ***rows*** Chocolate and Vanilla are closer to each other than to Strawberry because they have more similar profiles: they are high for the same columns and low for the same columns.

	Baseball	Basketball	Football	Soccer	Tennis
Chocolate	11	20	17	11	11
Vanilla	6	17	19	7	17
Strawberry	18	8	8	17	13

Correspondence Analysis: Example

Example:

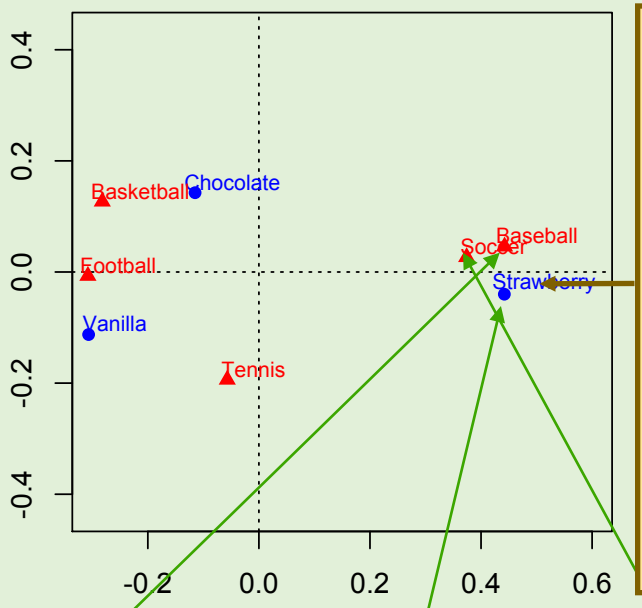


The ***columns*** Soccer and Baseball are close in the resulting plot because they have similar *profiles*: they are high for the same rows and low for the same rows.

	Baseball	Basketball	Football	Soccer	Tennis
Chocolate	11	20	17	11	11
Vanilla	6	17	19	7	17
Strawberry	18	8	8	17	13

Correspondence Analysis: Example

Example:



The ***columns*** Soccer and Baseball are close to the ***row*** Strawberry in the resulting plot because both Soccer and Baseball have higher counts with Strawberry than expected.

	Baseball	Basketball	Football	Soccer	Tennis
Chocolate	11	20	17	11	11
Vanilla	6	17	19	7	17
Strawberry	18	8	8	17	13

Correspondence Analysis: Summary

Correspondence analysis is a *descriptive* technique used to visualize and summarize relationships between categorical variables.

We have considered only two-way tables: that is, tables comparing two different categorical variables. Correspondence analysis may be extended to an arbitrary number of categorical variables.