FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
FALL 2015
CORRESPONDENCE ANALYSIS

- we are going to look at variants of PCA that apply to frequency data (CA), network data (HITS), and text data (LSA)
- we are also going to look at a variant of PCA called MDS that uses dissimilarity/distance matrix in place of covariance/inner product matrix
- a recap of the matrix decompositions that we have been using and a new one that we will introduce:

  SVD: for PCA, FA, HITS, LSI; $A \in \mathbb{R}^{n\times p}$,
$$A = U\Sigma V^{\mathsf{T}}, \qquad U^{\mathsf{T}}U = I_n, \qquad V^{\mathsf{T}}V = I_p$$
  EVD: for CCA, MDS; $A \in \mathbb{R}^{p\times p}$ symmetric,
$$X = Q\Lambda Q^{\mathsf{T}}, \qquad Q^{\mathsf{T}}Q = I_p$$
  GEVD: for LDA; $A \in \mathbb{R}^{p\times p}$ symmetric, $B \in \mathbb{R}^{p\times p}$ symmetric nonsingular,
$$B^{-1}A = Q\Lambda Q^{\mathsf{T}}, \qquad Q^{\mathsf{T}}Q = I_p$$
  GSVD: for CA; $A \in \mathbb{R}^{n\times p}$, $D_1 \in \mathbb{R}^{n\times n}$ diagonal nonsingular, $D_2 \in \mathbb{R}^{p\times p}$ diagonal nonsingular,
$$A = U\Sigma V^{\mathsf{T}}, \qquad U^{\mathsf{T}}D_1^{-1}U = I_n, \qquad V^{\mathsf{T}}D_2^{-1}V = I_p$$

1. Correspondence analysis

- in this case, the data matrix $X \in \mathbb{R}^{n\times p}$ is (part of) a *contingency table*
- a contingency table is essentially a table of *count* or *frequency* data
- an example is the Greenacre smokers data set

	none	light	medium	heavy	**row total**
senior managers	4	2	3	2	**11**
junior managers	4	3	7	4	**18**
senior employees	25	10	12	4	**51**
junior employees	18	24	33	13	**88**
secretaries	10	6	7	2	**25**
column total	**61**	**54**	**62**	**25**	**193**

- the bold faced row and column totals as well as the grand total can all be computed from the matrix
$$X = \begin{bmatrix} 4 & 2 & 3 & 2 \\ 4 & 3 & 7 & 4 \\ 25 & 10 & 12 & 4 \\ 18 & 24 & 33 & 13 \\ 10 & 6 & 7 & 2 \end{bmatrix} \in \mathbb{R}^{5\times 4}$$

Date: November 27, 2015, version 1.2. Comments, bug reports: `lekheng@galton.uchicago.edu`.

so this matrix in the middle forms the essence of a contingency table

- here $n = 5$ is the number of staff categories and $p = 4$ is the number of smoking categories — in a contingency table, the row and column categories are treated on equal footing
- more generally, for a data matrix $X = [x_{ij}] \in \mathbb{R}^{n\times p}$ that comes from a contingency table, x_{ij}'s are *frequencies*, i.e.,

$$x_{ij} = \text{number of observations in a sample that falls into row category } i \text{ and column category } j$$

- alternatively, there are people who prefer to normalize by the grand total before forming the data matrix
- for the smokers data set, this gives

$$X = \begin{bmatrix} \frac{4}{193} & \frac{2}{193} & \frac{3}{193} & \frac{2}{193} \\ \frac{4}{193} & \frac{3}{193} & \frac{7}{193} & \frac{4}{193} \\ \frac{25}{193} & \frac{10}{193} & \frac{12}{193} & \frac{4}{193} \\ \frac{18}{193} & \frac{24}{193} & \frac{33}{193} & \frac{13}{193} \\ \frac{10}{193} & \frac{6}{193} & \frac{7}{193} & \frac{2}{193} \end{bmatrix} \in \mathbb{R}^{5\times 4}$$

- here x_{ij}'s are *relative frequencies* instead of frequencies
- the only difference is that the result would differ by a constant so we would just stick to the frequency version in the following
- the most notations for the row, column, and grand totals are

$$\begin{aligned} x_{i\bullet} &= \sum_{j=1}^{p} x_{ij}, \qquad i = 1, \dots, n, \\ x_{\bullet j} &= \sum_{i=1}^{n} x_{ij}, \qquad j = 1, \dots, p, \\ x_{\bullet\bullet} &= \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} = \sum_{i=1}^{n} x_{i\bullet} = \sum_{j=1}^{p} x_{\bullet j} = \mathbf{1}_n^{\mathsf{T}} X \mathbf{1}_p \end{aligned}$$

- the objective in *correspondence analysis* (CA) is to find a *row weight vector* and a *column weight vector*,

$$\mathbf{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \in \mathbb{R}^n \qquad \text{and} \qquad \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} \in \mathbb{R}^p$$

such that

$$\begin{aligned} r_i &\propto \sum_{j=1}^{p} c_j \frac{x_{ij}}{x_{i\bullet}}, \qquad i = 1, \dots, n, \\ c_j &\propto \sum_{i=1}^{n} r_i \frac{x_{ij}}{x_{\bullet j}}, \qquad j = 1, \dots, p, \end{aligned}$$

hold simultaneously

- here '$x \propto y$' means 'x is proportional to y,' i.e., $x = \lambda y$ for some constant $\lambda \in \mathbb{R}$
- now let

$$D_r = \begin{bmatrix} x_{1\bullet} & & \\ & \ddots & \\ & & x_{n\bullet} \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad D_c = \begin{bmatrix} x_{\bullet 1} & & \\ & \ddots & \\ & & x_{\bullet p} \end{bmatrix} \in \mathbb{R}^{p\times p}$$

and rewrite the above equations in terms of vectors, we get

$$\mathbf{r} \propto D_r^{-1} X \mathbf{c}, \qquad \mathbf{c} \propto D_c^{-1} X^{\mathsf{T}} \mathbf{r} \tag{1.1}$$

- substituting one into the other, we get

$$\mathbf{r} \propto D_r^{-1} X D_c^{-1} X^{\mathsf{T}} \mathbf{r}, \qquad \mathbf{c} \propto D_c^{-1} X^{\mathsf{T}} D_r^{-1} X \mathbf{c} \tag{1.2}$$

- so $\mathbf{r} \in \mathbb{R}^n$ is an eigenvector of $D_r^{-1} X D_c^{-1} X^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ and $\mathbf{c} \in \mathbb{R}^p$ is an eigenvector of $D_c^{-1} X^{\mathsf{T}} D_r^{-1} X \in \mathbb{R}^{p\times p}$
- note that the nonzero eigenvalues of $D_r^{-1} X D_c^{-1} X^{\mathsf{T}} = D_r^{-1} X (X D_c^{-1})^{\mathsf{T}}$ and $D_c^{-1} X^{\mathsf{T}} D_r^{-1} X = (X D_c^{-1})^{\mathsf{T}} D_r^{-1} X$ are the same since the nonzero eigenvalues of AB always equals the nonzero eigenvalues of BA for any $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times n}$
- we are only interested in nonzero eigenvalues and so (1.2) becomes

$$D_r^{-1} X D_c^{-1} X^{\mathsf{T}} \mathbf{r} = \lambda \mathbf{r}, \qquad D_c^{-1} X^{\mathsf{T}} D_r^{-1} X \mathbf{c} = \lambda \mathbf{c}, \tag{1.3}$$

where $0 \neq \lambda \in \mathbb{R}$
- if $D_r = I_n$ and $D_c = I_p$, i.e., the $n \times n$ and $p \times p$ identity matrices, then (1.3) reduces to

$$X X^{\mathsf{T}} \mathbf{r} = \mathbf{r}, \qquad X^{\mathsf{T}} X \mathbf{c} = \mathbf{c},$$

i.e., $\mathbf{r}$ and $\mathbf{c}$ are left and right singular vectors of X

2. Generalized singular value decomposition

- but $D_r \neq I_n$ and $D_c \neq I_p$ in general and that's why we need the *generalized singular value decomposition* or GSVD

$$X = U \Sigma V^{\mathsf{T}}, \qquad U^{\mathsf{T}} D_r^{-1} U = I_n, \qquad V^{\mathsf{T}} D_c^{-1} V = I_p, \tag{2.1}$$

which can be shown to exist for any symmetric positive definite matrices $D_1 \in \mathbb{R}^{n\times n}$ and $D_2 \in \mathbb{R}^{p\times p}$ and any $X \in \mathbb{R}^{n\times p}$
- for our purpose we just need $D_r = \operatorname{diag}(x_{1\bullet}, \dots, x_{n\bullet})$ and $D_c = \operatorname{diag}(x_{\bullet 1}, \dots, x_{\bullet p})$, obviously symmetric positive definite since $x_{i\bullet} > 0$ and $x_{\bullet j} > 0$ for all $i = 1, \dots, n$ and $j = 1, \dots, p$
- we will call $\mathbf{u}_k \in \mathbb{R}^n$, $\mathbf{v}_k \in \mathbb{R}^p$ the left and right *generalized singular vectors* and $\sigma_k \in \mathbb{R}$ *generalized singular values* of $X \in \mathbb{R}^{n\times p}$ with respect to *weight matrices* D_r and D_c
- so you know how to compute GSVD or use a software that has a GSVD function, then you could compute the row and column weight vectors as generalized left and right singular vectors, i.e.,

$$\mathbf{r}_k = \mathbf{u}_k, \qquad \mathbf{c}_k = \mathbf{v}_k \tag{2.2}$$

- however GSVD is not as common as SVD, for instance it's not in R unless you load some additional packages
- in the following we will see how to do CA with the usual SVD

3. Correspondence analysis with SVD

- we need first deal with an issue: check that

$$D_r^{-1} X D_c^{-1} X^{\mathsf{T}} \mathbf{1}_n = \mathbf{1}_n, \qquad D_c^{-1} X^{\mathsf{T}} D_r^{-1} X \mathbf{1}_p = \mathbf{1}_p,$$

i.e., these matrices have a trivial eigenvector, a vector of all ones, corresponding to the eigenvalue 1, that we want to exclude from our possibilities for $\mathbf{r}$ and $\mathbf{c}$
- in other words, we want to exclude the left generalized singular vector $\mathbf{1}_n$ and right generalized singular vector $\mathbf{1}_p$ corresponding to the generalized singular value 1 from the GSVD of X

- how do you 'exclude' a pair of left/right singular vectors $\mathbf{u}_i, \mathbf{v}_i$, corresponding to a singular value σ_i? the answer is *deflation*, i.e., just subtract the rank-1 matrix $\sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ from X:

$$X - \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

- why does this work? remember that the SVD of X can be written as

$$X = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \cdots + \sigma_i \mathbf{u}_i \mathbf{v}_i^\top + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\top$$

  and if we subtract $\sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, we get

$$X - \sigma_i \mathbf{u}_i \mathbf{v}_i^\top = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \cdots + \sigma_{i-1} \mathbf{u}_{i-1} \mathbf{v}_{i-1}^\top + \sigma_{i+1} \mathbf{u}_{i+1} \mathbf{v}_{i+1}^\top + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\top,$$

  i.e., the term $\sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ no longer appears in the SVD of $X - \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$
- this also works for GSVD, but we have to scale the rank-1 term accordingly
- doing deflation gives us the matrix

$$X - \frac{\mathbf{a}\mathbf{b}^\top}{x_{\bullet\bullet}} \in \mathbb{R}^{n \times p} \tag{3.1}$$

  where

$$\mathbf{a} = D_r \mathbf{1}_n \in \mathbb{R}^n, \qquad \mathbf{b} = D_c \mathbf{1}_p \in \mathbb{R}^p$$

- now we could perform GSVD to (3.1) but we want to avoid GSVD and use only SVD, so we consider the matrix

$$Y := \sqrt{x_{\bullet\bullet}} D_r^{-1/2} \left(X - \frac{\mathbf{a}\mathbf{b}^\top}{x_{\bullet\bullet}} \right) D_c^{-1/2} \in \mathbb{R}^{n \times p}$$

  instead
- if we work out the entries of Y, we would see that

$$y_{ij} = \frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

  where

$$e_{ij} = \frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}$$

  for $i = 1, \dots, n$ and $j = 1, \dots, p$
- the square of the Frobenius norm of Y is then

$$\|Y\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p y_{ij}^2 = \sum_{i=1}^n \sum_{j=1}^p \frac{(x_{ij} - e_{ij})^2}{e_{ij}} \tag{3.2}$$

- for those who know some statistics the expression on the right of (3.2) is a χ^2-*test statistics* and e_{ij} may be interpreted as the *estimated expected value*
- now we may perform a usual SVD to Y to obtain

$$Y = U \Sigma V^\top, \qquad U^\top U = I_n, \qquad V^\top V = I_p$$

  where $U = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ and $V = [\mathbf{v}_1, \dots, \mathbf{v}_p]$
- in particular

$$Y \mathbf{v}_k = \sigma_k \mathbf{u}_k, \qquad Y^\top \mathbf{u}_k = \sigma_k \mathbf{v}_k, \qquad k = 1, \dots, r$$

  where $r = \operatorname{rank}(Y)$
- we claim that defining $\mathbf{r}_k := \sigma_k D_r^{-1/2} \mathbf{u}_k$ and $\mathbf{c}_k := \sigma_k D_c^{-1/2} \mathbf{v}_k$ gives us a solution to (1.1)

- to show this, note that
$$\begin{aligned}\mathbf{r}_k &= \sigma_k D_r^{-1/2}\mathbf{u}_k = D_r^{-1/2}(\sigma_k \mathbf{u}_k) = D_r^{-1/2} Y \mathbf{v}_k \\ &= D_r^{-1/2} Y \left(\frac{1}{\sigma_k} D_c^{1/2}\mathbf{c}_k\right) = \frac{1}{\sigma_k} D_r^{-1/2} Y D_c^{1/2}\mathbf{c}_k,\end{aligned}$$
and likewise
$$\mathbf{c}_k = \frac{1}{\sigma_k} D_c^{-1/2} Y^{\mathsf{T}} D_r^{1/2}\mathbf{r}_k$$
- since $\mathbf{a}^{\mathsf{T}}\mathbf{r}_k = 0$ and $\mathbf{b}^{\mathsf{T}}\mathbf{c}_k = 0$ (see appendix below), we get
$$\mathbf{r}_k = \frac{\sqrt{x_{\bullet\bullet}}}{\sigma_k} D_r^{-1} X \mathbf{c}_k, \qquad \mathbf{c}_k = \frac{\sqrt{x_{\bullet\bullet}}}{\sigma_k} D_c^{-1} X^{\mathsf{T}} \mathbf{r}_k \tag{3.3}$$
as required
- note that
$$\mathbf{r}_i^{\mathsf{T}} D_r \mathbf{r}_j = \sigma_i\sigma_j \mathbf{u}_i^{\mathsf{T}} D_r^{-1/2} D_r D_r^{-1/2}\mathbf{u}_j = \sigma_i\sigma_j \mathbf{u}_i^{\mathsf{T}}\mathbf{u}_j = \begin{cases}\sigma_i^2 & \text{if } i = j,\\ 0 & \text{if } i \neq j,\end{cases}$$
and
$$\mathbf{c}_i^{\mathsf{T}} D_c \mathbf{c}_j = \sigma_i\sigma_j \mathbf{v}_i^{\mathsf{T}} D_c^{-1/2} D_c D_c^{-1/2}\mathbf{v}_j = \sigma_i\sigma_j \mathbf{v}_i^{\mathsf{T}}\mathbf{v}_j = \begin{cases}\sigma_i^2 & \text{if } i = j,\\ 0 & \text{if } i \neq j,\end{cases}$$
- this is not surprising since the method using GSVD (2.2) and the method using SVD give the same solution (3.3) up to a constant
- the sample means of $\mathbf{r}_k$ and $\mathbf{c}_k$ are both zero since
$$\frac{1}{x_{\bullet\bullet}}\mathbf{a}^{\mathsf{T}}\mathbf{r}_k = 0, \qquad \frac{1}{x_{\bullet\bullet}}\mathbf{b}^{\mathsf{T}}\mathbf{c}_k = 0$$
- the sample variances of $\mathbf{r}_k$ and $\mathbf{c}_k$ are both $\sigma_k^2/x_{\bullet\bullet}$ since
$$\frac{1}{x_{\bullet\bullet}}\sum_{i=1}^{n} x_{i\bullet} r_{ki}^2 = \frac{\mathbf{r}_k^{\mathsf{T}} D_r \mathbf{r}_k}{x_{\bullet\bullet}} = \frac{\sigma_k^2}{x_{\bullet\bullet}}$$
and
$$\frac{1}{x_{\bullet\bullet}}\sum_{j=1}^{p} x_{\bullet j} c_{kj}^2 = \frac{\mathbf{c}_k^{\mathsf{T}} D_c \mathbf{c}_k}{x_{\bullet\bullet}} = \frac{\sigma_k^2}{x_{\bullet\bullet}}$$
- we use CA the way we use PCA — scree plots, scatter plots, biplots — but with $\mathbf{r}_k$ and $\mathbf{c}_k$ playing the roles of $\mathbf{u}_k$ and $\mathbf{v}_k$ in PCA

4. APPENDIX

- we verify that $\mathbf{a}^{\mathsf{T}}\mathbf{r}_k = 0$ and $\mathbf{b}^{\mathsf{T}}\mathbf{c}_k = 0$
- for your convenience
$$\mathbf{r}_k = \sigma_k D_r^{-1/2}\mathbf{u}_k, \qquad \mathbf{a} = D_r \mathbf{1}_n, \qquad \mathbf{b} = D_c \mathbf{1}_p,$$
$$Y = \sqrt{x_{\bullet\bullet}} D_r^{-1/2}\left(X - \frac{\mathbf{a}\mathbf{b}^{\mathsf{T}}}{x_{\bullet\bullet}}\right) D_c^{-1/2}, \qquad Y\mathbf{v}_k = \sigma_k \mathbf{u}_k$$
- so
$$\begin{aligned}\mathbf{a}^{\mathsf{T}}\mathbf{r}_k &= \mathbf{1}_n^{\mathsf{T}} D_r^{1/2}(\sigma_k \mathbf{u}_k) = \mathbf{1}_n^{\mathsf{T}} D_r^{1/2} Y \mathbf{v}_k \\ &= \sqrt{x_{\bullet\bullet}}\mathbf{1}_n^{\mathsf{T}}\left(X - \frac{\mathbf{a}\mathbf{b}^{\mathsf{T}}}{x_{\bullet\bullet}}\right) D_c^{-1/2}\mathbf{v}_k\end{aligned}$$

- we claim that
$$\left(X - \frac{\mathbf{a}\mathbf{b}^\mathsf{T}}{x_{\bullet\bullet}}\right)^\mathsf{T} \mathbf{1}_n = \mathbf{0}_p$$
- this follows from
$$X^\mathsf{T}\mathbf{1}_n = \begin{bmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ \vdots \\ x_{\bullet p} \end{bmatrix}$$
and
$$\mathbf{b}\mathbf{a}^\mathsf{T}\mathbf{1}_n = (\mathbf{1}_n^\mathsf{T} D_r \mathbf{1}_n)\mathbf{b} = (x_{1\bullet} + \cdots + x_{n\bullet})\mathbf{b} = x_{\bullet\bullet}\mathbf{b}$$
and so
$$\left(\frac{\mathbf{a}\mathbf{b}^\mathsf{T}}{x_{\bullet\bullet}}\right)^\mathsf{T} \mathbf{1}_n = \mathbf{b} = D_c\mathbf{1}_p = \begin{bmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ \vdots \\ x_{\bullet p} \end{bmatrix}$$