Distances and similarities
Based in part on slides from textbook, slides of Susan Holmes

October 3, 2012
Distances and similarities

**Similarities**

- Start with $X$ which we assume is centered and standardized.
- The PCA loadings were given by eigenvectors of the *correlation matrix* which is a measure of similarity.
- The first 2 (or any 2) PCA scores yield an $n \times 2$ matrix that can be visualized as a scatter plot.
- Similarly, the first 2 (or any 2) PCA loadings yield an $p \times 2$ matrix that can be visualized as a scatter plot.
Olympic data
Distances and similarities

**Similarities**

- The PCA loadings were given by eigenvectors of the *correlation matrix* which is a measure of *similarity*.
- The visualization of the cases based on PCA scores were determined by the eigenvectors of $\mathbf{XX}^T$, an $n \times n$ matrix.
- To see this, remember that

$$\mathbf{X} = \mathbf{U}\Delta \mathbf{V}^T$$

so

$$\mathbf{XX}^T = \mathbf{U}\Delta^2 \mathbf{U}^T$$
$$\mathbf{X}^T \mathbf{X} = \mathbf{V}\Delta^2 \mathbf{V}^T$$
Distances and similarities

Similarities

- The matrix $XX^T$ is a measure of similarity between cases.
- The matrix $X^TX$ is a measure of similarity between features.
- Structure in the two similarity matrices yield insight into the set of cases, or the set of features . . .
Distances and similarities

Distances

- Distances are inversely related to similarities.
- If $A$ and $B$ are similar, then $d(A, B)$ should be small, i.e. they should be near.
- If $A$ and $B$ are distant, then they should not be similar.
- For a data matrix, there is a natural distance between cases:

$$d(X_i, X_k)^2 = \|X_i - X_k\|_2^2$$

$$= \sum_{j=1}^{p} \|X_{ij} - X_{kj}\|_2^2$$

$$= (XX^T)_{ii} - 2(XX^T)_{ik} + (XX^T)_{kk}$$
Distances and similarities

Distances

- Suggests a natural transformation between a similarity matrix $S$ and a distance matrix $D$

$$D_{ik} = (S_{ii} - 2 \cdot S_{ik} + S_{kk})^{1/2}$$

- The reverse transformation is not so obvious. Some suggestions from your book:

$$S_{ik} = -D_{ik}$$

$$= e^{-D_{ik}}$$

$$= \frac{1}{1 + D_{ik}}$$
Distances and similarities

Distances

- A distance (or a metric) on a set $S$ is a function $d : S \times S \rightarrow [0, +\infty)$ that satisfies
  - $d(x, x) = 0; d(x, y) = 0 \iff x = y$
  - $d(x, y) = d(y, x)$
  - $d(x, y) \leq d(x, z) + d(z, y)$

- If $d(x, y) = 0$ for some $x \neq y$ then $d$ is a pseudo-metric.
Distances and similarities

Similarities

- A similarity on a set $S$ is a function $s : S \times S \rightarrow \mathbb{R}$ and should satisfy
  - $s(x, x) \geq s(x, y)$ for all $x \neq y$
  - $s(x, y) = s(y, x)$

- By adding a constant, we can often assume that $s(x, y) \geq 0$. 
Distances and similarities

Examples: nominal data

• The simplest example for nominal data is just the discrete metric

\[ d(x, y) = \begin{cases} 0 & x = y \\ 1 & \text{otherwise} \end{cases} \]

• The corresponding similarity would be

\[ s(x, y) = \begin{cases} 1 & x = y \\ 0 & \text{otherwise} \end{cases} \]

\[ = 1 - d(x, y) \]
Distances and similarities

Examples: ordinal data

- If $S$ is ordered, we can think of $S$ as (or identify $S$ with) a subset of the non-negative integers.
- If $|S| = m$ then a natural distance is
  \[
  d(x, y) = \frac{|x - y|}{m - 1} \leq 1
  \]
- The corresponding similarity would be
  \[
  s(x, y) = 1 - d(x, y)
  \]
Distances and similarities

Examples: vectors of continuous data

- If $S = \mathbb{R}^k$ there are lots of distances determined by norms.
- The Minkowski $p$ or $\ell_p$ norm, for $p \geq 1$:

$$d(x, y) = \|x - y\|_p = \left( \sum_{i=1}^{k} |x_i - y_i|^p \right)^{1/p}$$

- Examples:
  - $p = 2$ the usual Euclidean distance, $\ell_2$
  - $p = 1$ $d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$, the “taxicab distance”, $\ell_1$
  - $p = \infty$ the $d(x, y) = \max_{1 \leq i \leq k} |x_i - y_i|$, the sup norm, $\ell_{\infty}$
Distances and similarities

Examples: vectors of continuous data

- If $0 < p < 1$ this is not a norm, but it still defines a metric.
- The preceding transformations can be used to construct similarities.

Examples: binary vectors

- If $S = \{0, 1\}^k \subset \mathbb{R}^k$ the vectors can be thought of as vectors of bits.
- The $\ell_1$ norm counts the number of mismatched bits.
- This is known as Hamming distance.
Distances and similarities

Example: Mahalanobis distance

- Given $\Sigma_{k\times k}$ that is positive definite, we define *Mahalanobis distance* on $\mathbb{R}^k$ by

  $$d_\Sigma(x, y) = \left( (x - y)^T \Sigma^{-1} (x - y) \right)^{1/2}$$

- This is the usual Euclidean distance, with a change of basis given by a rotation and stretching of the axes.

- If $\Sigma$ is only non-negative definite, then we can replace $\Sigma^{-1}$ with $\Sigma^\dagger$, its *pseudo-inverse*. This yields a *pseudo-metric* because it fails the test

  $$d(x, y) = 0 \iff x = y.$$
Distances and similarities

### Example: similarities for binary vectors

- Given binary vectors $x, y \in \{0, 1\}^k$ we can summarize their agreement in a $2 \times 2$ table:

<table>
<thead>
<tr>
<th></th>
<th>$x = 0$</th>
<th>$x = 1$</th>
<th>total$_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = 0$</td>
<td>$f_{00}$</td>
<td>$f_{01}$</td>
<td>$f_0$</td>
</tr>
<tr>
<td>$y = 1$</td>
<td>$f_{10}$</td>
<td>$f_{11}$</td>
<td>$f_1$</td>
</tr>
<tr>
<td>total$_x$</td>
<td>$f_0$</td>
<td>$f_1$</td>
<td>$k$</td>
</tr>
</tbody>
</table>
Distances and similarities

Example: similarities for binary vectors

- We define the simple matching coefficient (SMC) similarity by

\[ SMC(x, y) = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{f_{00} + f_{11}}{k} \]

\[ = 1 - \frac{\|x - y\|_1}{k} \]

- The Jaccard coefficient ignores entries where \( x_i = y_i = 0 \)

\[ J(x, y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \]
Distances and similarities

Example: cosine similarity & correlation

- Given vectors $x, y \in \mathbb{R}^k$ the cosine similarity is defined as

$$-1 \leq \cos(x, y) = \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} \leq 1.$$ 

- The correlation between two vectors is defined as

$$-1 \leq \text{cor}(x, y) = \frac{\langle x - \bar{x} \cdot 1, y - \bar{y} \cdot 1 \rangle}{\|x - \bar{x} \cdot 1\|_2 \|y - \bar{y} \cdot 1\|_2} \leq 1.$$
Distances and similarities

Example: correlation

- An alternative, perhaps more familiar definition:

\[
\text{cor}(x, y) = \frac{S_{xy}}{S_x S_y}
\]

\[
S_{xy} = \frac{1}{k - 1} \sum_{i=1}^{k} (x_i - \bar{x})(y_i - \bar{y})
\]

\[
S_x^2 = \frac{1}{k - 1} \sum_{i=1}^{k} (x_i - \bar{x})^2
\]

\[
S_y^2 = \frac{1}{k - 1} \sum_{i=1}^{k} (y_i - \bar{y})^2
\]
Distances and similarities

Correlation & PCA

- The matrix $\frac{1}{n-1} \tilde{X}^T \tilde{X}$ was actually the matrix of pair-wise correlations of the features. Why? How?

\[
\frac{1}{n-1} (\tilde{X}^T \tilde{X})_{ij} = \frac{1}{n-1} \left( D^{-1/2} X^T H X D^{-1/2} \right)
\]

- The diagonal entries of $D$ are the sample variances of each feature.

- The inner matrix multiplication computes the pair-wise dot-products of the columns of $HX$

\[
(X^T H X)_{ij} = \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)
\]
High positive correlation
High negative correlation
No correlation
Small positive
Small negative
Distances and similarities

In a given data set, each case may have many attributes or features.

Example: see the health data set for HW 1.

To compute similarities of cases, we must pool similarities across features.

In a data set with $M$ different features, we write $x_i = (x_{i1}, \ldots, x_{iM})$, with each $x_{im} \in S_m$. 

Combining similarities
Distances and similarities

Combining similarities

- Given similarities $s_m$ on each $S_m$ we can define an overall similarity between case $x_i$ and $x_j$ by

$$s(x_i, x_j) = \sum_{m=1}^{M} w_m s_m(x_{im}, x_{jm})$$

with optional weights $w_m$ for each feature.

- Your book modifies this to deal with “asymmetric attributes”, i.e. attributes for which Jaccard similarity might be used.