Other datatypes

Document data

- You might start with a collection of $n$ documents (i.e. all of Shakespeare’s works in some digital format; all of Wikileaks’ U.S. Embassy Cables).
- This is not a data matrix . . .
- Given $p$ terms of interest: $\{\text{Al Qaeda, Iran, Iraq, etc.}\}$ one can form a term-document matrix filled with counts.
Time series

- Imagine recording the minute-by-minute prices of all stocks in the S & P 500 for last 200 days of trading.
- The data can be represented by a $78000 \times 500$ matrix.
- BUT, there is definite structure across the rows of this matrix.
- They are not unrelated “cases” like they might be in other applications.

Transaction data

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Social media data

- Bruce Arthur [Link](https://twitter.com/BruceArthur/status/1234567890)

Graph data

![Graph diagram]
Other data types

Graph data
- Nodes on the graph might be Facebook users, or public pages.
- Weights on the edges could be number of messages sent in a prespecified period.
- If you take weekly intervals, this leads to a sequence of graphs

\[ G_i = \text{communication over } i\text{-th week.} \]
- How this graph changes is of obvious interest . . .
- Even structure of just one graph is of interest – we’ll come back to this when we talk about spectral methods . . .

Data quality

Some issues to keep in mind
- Is the data experimental or observational?
- If observational, what do we know about the data generating mechanism? For example, although the S&P 500 example can be represented as a data matrix, there is clearly structure across rows.
- General quality issues:
  - How much of the data missing? Is missingness informative?
  - Is it very noisy? Are there outliers?
  - Are there a large number of duplicates?

Preprocessing

General procedures
- Aggregation Combining features into a new feature. Example: pooling county-level data to state-level data.
- Discretization Breaking up a continuous variable (or a set of continuous variables) into an ordinal (or nominal) discrete variable.
- Transformation Simple transformation of feature (log or exp) or mapping to a new space (Fourier transform / power spectrum, wavelet transform).

A continuous variable that could be discretized
Discretization by fixed width

Discretization by fixed quartile

Discretization by clustering

Variable transformation: bacterial decay
Variable transformation: bacterial decay

BMW daily returns (fEcofin package)

ACF of BMW daily returns

Discrete wavelet transform of BMW daily returns
Part II

Dimension reduction, PCA &
eigenanalysis

Dimension reduction

- By choosing $\beta$ appropriately, we may find "interesting" new features.
- Suppose we take $k$ much smaller than $p$ vectors of $\beta$ which we write as a matrix $B_{p \times k}$.
- The new data matrix $XB$ has fewer dimensions than $X$.
- This is dimension reduction ...

Combinations of features

- Given a data matrix $X_{n \times p}$ with $p$ fairly large, it can be difficult to visualize structure.
- Often useful to look at linear combinations of the features.
- Each $\beta \in \mathbb{R}^p$, determines a linear rule
  $$f_\beta(x) = x^T \beta$$
- Evaluating this on each row $X_i$ of $X$ yields a vector
  $$(f_\beta(X_i))_{1 \leq i \leq n} = X\beta.$$
Dimension reduction

Principal Components
- Define the $n \times n$ matrix
  $$H = I_{n \times n} - \frac{1}{n}11^T$$
- This matrix removes means:
  $$(Hv)_i = v_i - \bar{v}.$$ 
- It is also a projection matrix:
  $$H^T = H$$
  $$H^2 = H$$

Dimension reduction

Eigenanalysis
- The matrix $X^T H X$ is symmetric, so it can be written as
  $$X^T H X = V D V^T$$
  where $D_{k \times k} = \text{diag}(d_1,\ldots,d_k)$, $\text{rank}(X^T H X) = k$ and
  $V_{p \times k}$ has orthonormal columns, i.e. $V^T V = I_{k \times k}$.
- We always have $d_i \geq 0$ and we take $d_1 \geq d_2 \geq \ldots d_k$.

Dimension reduction

Principal Components with Matrices
- With this matrix,
  $$\hat{\text{Var}}(V \beta) = \frac{1}{n - 1} \beta^T X^T H X \beta.$$ 
- So, maximizing sample variance, with $\|\beta\|_2 = 1$ is
  $$\text{maximize } \beta^T \left( X^T H X \right) \beta, \quad \|\beta\|_2 = 1$$
  $$\text{This boils down to an eigenvalue problem . . .}$$

Dimension reduction

Eigenanalysis & PCA
- Suppose now that $\beta = \sum_{j=1}^k a_j v_j$ with $v_j$ the columns of $V$. Then,
  $$\|\beta\|_2 = \sqrt{\sum_{i=1}^k a_i^2}$$
  $$\beta^T \left( X^T H X \right) \beta = \sum_{i=1}^k a_i^2 d_i$$
- Choosing $a_1 = 1$, $a_j = 0, j \geq 2$ maximizes this quantity.
Dimension reduction

Eigenanalysis & PCA

- Therefore, \( \hat{\beta}_1 = v_1 \) the first column of \( V \) solves
  \[
  \maximize \beta^T \left( X^T H X \right) \beta, \quad \|\beta\|_2 = 1
  \]

- This yields scores \( HX\hat{\beta}_1 \in \mathbb{R}^n \).
- These are the 1st principal component scores.

Higher order components

- In matrix terms, all the principal components scores are
  \[
  (HXV)_{n \times k}
  \]
  and the loadings are the columns of \( V \).
- This information can be summarized in a biplot.
- The loadings describe how each feature contributes to each principal component score.

Dimension reduction

Higher order components

- Having found the direction with “maximal sample variance” we might look for the “second most variable” direction by solving
  \[
  \maximize \beta^T \left( X^T H X \right) \beta, \quad \|\beta\|_2 = 1, \beta^T v_1 = 0
  \]

- Note we restricted our search so we would not just recover \( v_1 \) again.
- Not hard to see that if \( \beta = \sum_{j=1}^k a_j v_j \) this is solved by taking \( a_2 = 1, a_j = 0, j \neq 2 \).

Olympic data

In matrix terms, all the principal components scores are

\[
(HXV)_{n \times k}
\]

and the loadings are the columns of \( V \).
This information can be summarized in a biplot.
The loadings describe how each feature contributes to each principal component score.
Dimension reduction

The importance of scale

- The PCA scores are not invariant to scaling of the features: for $Q_{p \times p}$ diagonal the PCA scores of $XQ$ are not the same as $X$.
- Common to convert all variables to the same scale before applying PCA.
- Define the scalings to be the sample standard deviation of each feature. In matrix form, let

$$ S^2 = \frac{1}{n-1} \text{diag} \left( X^T H X \right) $$

- Define $\tilde{X} = HX S^{-1}$. The normalized PCA loadings are given by an eigenanalysis of $\tilde{X}^T \tilde{X}$. 

Olympic data: screeplot
**Dimension reduction**

**PCA and the SVD**

- The singular value decomposition of a matrix tells us we can write

\[ \tilde{X} = U \Delta V^T \]

with \( \Delta_{k \times k} = \text{diag}(\delta_1, \ldots, \delta_k) \), \( \delta_j \geq 0 \), \( k = \text{rank}(\tilde{X}) \), \( U^T U = V^T V = I_{k \times k} \).

- Recall that the scores were

\[ \tilde{X} V = (U \Delta V^T) V = U \Delta \]

- Also,

\[ \tilde{X}^T \tilde{X} = V \Delta^2 V^T \]

so \( D = \Delta^2 \).

**Another characterization of SVD**

- Given a data matrix \( X \) (or its scaled centered version \( \tilde{X} \)) we might try solving

\[ \min_{Z: \text{rank}(Z) = k} \|X - Z\|_F^2 \]

where \( F \) stands for Frobenius norm on matrices

\[ \|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2 \]

- It can be proven that

\[ Z = S_{n \times k} (V^T)_{k \times p} \]

where \( S_{n \times k} \) is the matrix of the first \( k \) PCA scores and the columns of \( V \) are the first \( k \) PCA loadings.

- This approximation is related to the screeplot: the height of each bar describes the additional drop in Frobenius norm as the rank of the approximation increases.
Dimension reduction

Other types of dimension reduction

- Instead of maximizing sample variance, we might try maximizing some other quantity . . .
- Independent Component Analysis tries to maximize “non-Gaussianity” of $V_\beta$. In practice, it uses skewness, kurtosis or other moments to quantify “non-Gaussianity.”
- These are both unsupervised approaches.
- Often, these are combined with supervised approaches into an algorithm like:
  - Feature creation: Build some set of features using PCA.
  - Validate: Use the derived features to see if they are helpful in the supervised problem.

Olympic data: 1st component vs. total score
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.

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
- 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
- A distance (or a metric) on a set $S$ is a function $d : S \times S \to [0, +\infty)$ 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.

Similarities
- A similarity on a set $S$ is a function $s : S \times S \to \mathbb{R}$ 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
  \[ s(x, y) = \begin{cases} 
    1 & x = y \\
    0 & \text{otherwise.} 
  \end{cases} \]
  \[ = 1 - d(x, y) \]

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
  \[ s(x, y) = 1 - d(x, y) \]

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, \( \ell_2 \)
  - \( p = 1 \) the “taxicab distance”, \( \ell_1 \)
  - \( p = \infty \) the sup norm, \( \ell_\infty \)

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 the 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

- 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: correlation

- An alternative, perhaps more familiar definition:

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

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

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

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

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?

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

- 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 \(H X\)

\[
\left( X^T H X \right)_{ij} = \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)
\]

High positive correlation

High negative correlation
No correlation

Small positive

Small negative

Distances and similarities

Combining 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

- 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.