A few applications of the SVD

Many methods require to approximate the original data (matrix) by a low rank matrix before attempting to solve the original problem.

- Regularization methods require the solution of a least-squares linear system $Ax = b$ approximately in the dominant singular space of $A$.

- The Latent Semantic Indexing (LSI) method in information retrieval, performs the “query” in the dominant singular space of $A$.

- Methods utilizing Principal Component Analysis, e.g. Face Recognition.
**Commonality:** Approximate $\mathbf{A}$ (or $\mathbf{A}^\dagger$) by a lower rank approximation $\mathbf{A}_k$ (using dominant singular space) before solving original problem.

- This approximation captures the main features of the data while getting rid of noise and redundancy.

**Note:** Common misconception: ‘we need to reduce dimension in order to reduce computational cost’. In reality: using less information often yields better results. This is the problem of overfitting.

- Good illustration: Information Retrieval (IR)
Given: a collection of documents (columns of a matrix $A$) and a query vector $q$.

Collection represented by an $m \times n$ term by document matrix with $a_{ij} = L_{ij}G_iN_j$

Queries (‘pseudo-documents’) $q$ are represented similarly to a column

Information Retrieval: Vector Space Model
Vector Space Model - continued

- Problem: find a column of \( A \) that best matches \( q \)
- Similarity metric: angle between the column and \( q \) - Use cosines:

\[
\frac{|c^T q|}{\|c\|_2 \|q\|_2}
\]

- To rank all documents we need to compute

\[
s = A^T q
\]

- \( s = \) similarity vector.
- Literal matching – not very effective.
Use of the SVD

Many problems with literal matching: polysemy, synonymy, ...

Need to extract intrinsic information – or underlying “semantic” information –

Solution (LSI): replace matrix $A$ by a low rank approximation using the Singular Value Decomposition (SVD)

$$A = U \Sigma V^T \rightarrow A_k = U_k \Sigma_k V_k^T$$

- $U_k$: term space, $V_k$: document space.

- Refer to this as Truncated SVD (TSVD) approach
New similarity vector:

\[ s_k = A_k^T q = V_k \Sigma_k U_k^T q \]

**Issues:**

- Problem 1: How to select \( k \)?
- Problem 2: computational cost (memory + computation)
- Problem 3: updates [e.g. google data changes all the time]
- Not practical for very large sets
LSI : an example

%% D1 : INFANT & TODDLER first aid
%% D2 : BABIES & CHILDREN’s room for your HOME
%% D3 : CHILD SAFETY at HOME
%% D4 : Your BABY’s HEALTH and SAFETY
%% D5 : From INFANT to TODDLER
%% D6 : BABY PROOFING basics
%% D7 : Your GUIDE to easy rust PROOFING
%% D8 : Beanie BABIES collector’s GUIDE
%% D9 : SAFETY GUIDE for CHILD PROOFING your HOME
%% 6: INFANT 7: PROOFING 8: SAFETY 9: TODDLER
%% Source: Berry and Browne, SIAM., ’99

➤ Number of documents: 8
➤ Number of terms: 9
Raw matrix (before scaling).

\[
A = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & \text{bab} \\
1 & 1 & 1 & 1 & 1 & \text{chi} \\
1 & 1 & 1 & 1 & 1 & \text{gui} \\
1 & 1 & 1 & 1 & 1 & \text{hea} \\
1 & 1 & 1 & 1 & 1 & \text{hom} \\
1 & 1 & 1 & 1 & 1 & \text{inf} \\
1 & 1 & 1 & 1 & 1 & \text{pro} \\
1 & 1 & 1 & 1 & 1 & \text{saf} \\
1 & 1 & 1 & 1 & 1 & \text{tod}
\end{pmatrix}
\]

Get the answer to the query Child Safety, so

\[
q = [0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0]
\]

using cosines and then using LSI with \( k = 3 \).
Dimension reduction

Dimensionality Reduction (DR) techniques pervasive to many applications

- Often main goal of dimension reduction is not to reduce computational cost. Instead:
  - Dimension reduction used to reduce noise and redundancy in data
  - Dimension reduction used to discover patterns (e.g., supervised learning)

- Techniques depend on desirable features or application: Preserve angles? Preserve distances? Maximize variance? ..
The problem

Given $d \ll m$ find a mapping

$$\Phi : x \in \mathbb{R}^m \rightarrow y \in \mathbb{R}^d$$

- Mapping may be explicit (e.g., $y = V^T x$)
- Or implicit (nonlinear)

Practically:
Find a low-dimensional representation $Y \in \mathbb{R}^{d\times n}$ of $X \in \mathbb{R}^{m\times n}$.

- Two classes of methods: (1) projection techniques and (2) nonlinear implicit methods.
Example: Digit images (a sample of 30)
A few 2-D 'reductions':

PCA - digits: 0 -- 4

LLE - digits: 0 -- 4

K-PCA - digits: 0 -- 4

ONPP - digits: 0 -- 4

(articles) – SVDapp
**Projection-based Dimensionality Reduction**

*Given:* a data set \( X = [x_1, x_2, \ldots, x_n] \), and \( d \) the dimension of the desired reduced space \( Y \).

*Want:* a linear transformation from \( X \) to \( Y \)

\[
V^\top \in \mathbb{R}^{m \times d}
\]

\[
Y = V^\top X
\]

\[
\rightarrow Y \in \mathbb{R}^{d \times n}
\]

\( m \)-dimens. objects \((x_i)\) ‘flattened’ to \( d\)-dimens. space \((y_i)\)

*Problem:* Find the best such mapping (optimization) given that the \( y_i \)’s must satisfy certain constraints
Principal Component Analysis (PCA)

- PCA: find $V$ (orthogonal) so that projected data $Y = V^T X$ has maximum variance
- Maximize over all orthogonal $m \times d$ matrices $V$:

$$\sum_i \|y_i - \frac{1}{n} \sum_j y_j\|_2^2 = \cdots = \text{Tr} \left[ V^T \bar{X} \bar{X}^T V \right]$$

Where: $\bar{X} = [\bar{x}_1, \cdots, \bar{x}_n]$ with $\bar{x}_i = x_i - \mu$, $\mu = \text{mean}$.

Solution:

$V = \{ \text{dominant eigenvectors} \}$ of the covariance matrix

- i.e., Optimal $V = \text{Set of left singular vectors of } \bar{X} \text{ associated with } d \text{ largest singular values}$. 

11-14 (articles) – SVDapp
Show that $\bar{X} = X(I - \frac{1}{n}ee^T)$ (here $e = \text{vector of all ones}$). What does the projector $(I - \frac{1}{n}ee^T)$ do?

Show that solution $V$ also minimizes ‘reconstruction error’ ..

$$\sum_i \|\bar{x}_i - VV^T\bar{x}_i\|^2 = \sum_i \|\bar{x}_i - V\bar{y}_i\|^2$$

.. and that it also maximizes $\sum_{i,j} \|y_i - y_j\|^2$
Matrix Completion Problem

Consider a table of movie ratings. You want to predict missing ratings by assuming commonality (low rank matrix).

<table>
<thead>
<tr>
<th>movie</th>
<th>Paul</th>
<th>Jane</th>
<th>Ann</th>
<th>given data</th>
<th>predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title-1</td>
<td>−1</td>
<td>3</td>
<td>−1</td>
<td>−1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Title-2</td>
<td>4</td>
<td>x</td>
<td>3</td>
<td>2.8</td>
<td>−1.2</td>
</tr>
<tr>
<td>Title-3</td>
<td>−3</td>
<td>1</td>
<td>−4</td>
<td>−2.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Title-4</td>
<td>x</td>
<td>−1</td>
<td>−1</td>
<td>−0.5</td>
<td>−0.3</td>
</tr>
<tr>
<td>Title-5</td>
<td>3</td>
<td>−2</td>
<td>1</td>
<td>1.8</td>
<td>−1.4</td>
</tr>
<tr>
<td>Title-6</td>
<td>−2</td>
<td>3</td>
<td>x</td>
<td>−1.6</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Minimize \( \|(X - A)_{\text{mask}}\|_F^2 + \mu\|X\|_\ast \)

“minimize sum-of-squares of deviations from known ratings plus sum of singular values of solution (to reduce the rank).”