LECTURE 5
Singular Value Decomposition

The license plate of Gene Golub (1932–2007).

Outline

1. The singular value decomposition (SVD)
   - Motivation
   - Dyads
   - SVD theorem

2. Matrix properties via SVD
   - Rank, nullspace and range
   - Matrix norms, pseudo-inverses
   - Projectors

3. Low-rank matrix approximation

4. Principal component analysis (PCA)

Motivation

Figure: Votes of US Senators, 2002-2004. The plot is impossible to read...

- Can we make sense of this data?
- For example, do the two political parties emerge from the data alone?
- Could we further provide a "political spectrum score" for each Senator?
Dyads

A matrix $A \in \mathbb{R}^{m \times n}$ is called a dyad if it can be written as

$$A = pq^\top$$

for some vectors $p \in \mathbb{R}^m$, $q \in \mathbb{R}^n$. Element-wise the above reads

$$A_{ij} = p_i q_j, \quad 1 \leq i \leq m, \ 1 \leq j \leq n.$$  

Interpretation:
- The columns of $A$ are scaled copies of the same column $p$, with scaling factors given in vector $q$.
- The rows of $A$ are scaled copies of the same row $q^\top$, with scaling factors given in vector $p$.

**Example: video frames**

We are given a set of image frames representing a video. Assuming that each image is represented by a row vector of pixels, we can represent the whole video sequence as a matrix $A$. Each row is an image.

If the video shows a scene where no movement occurs, then the matrix $A$ is a dyad.

**Sums of dyads**

The singular value decomposition theorem, seen next, states that any matrix can be written as a sum of dyads:

$$A = \sum_{i=1}^{r} p_i q_i^\top$$

for vectors $p_i, q_i$ that are mutually orthogonal.

This allows to interpret data matrices as sums of “simpler” matrices (dyads).

**The singular value decomposition (SVD)**

The singular value decomposition (SVD) of a matrix provides a three-term factorization which is similar to the spectral factorization, but holds for any, possibly non-symmetric and rectangular, matrix $A \in \mathbb{R}^{m \times n}$.

**Theorem 1 (SVD decomposition)**

Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as

$$A = U\Sigma V^\top$$

where $V \in \mathbb{R}^{n \times n}$ and $U \in \mathbb{R}^{m \times m}$ are orthogonal matrices (i.e., $U^\top U = I_m$, $V^\top V = I_n$), and $\Sigma \in \mathbb{R}^{m \times n}$ is a matrix having the first $r = \text{rank}(A)$ diagonal entries $(\sigma_1, \ldots, \sigma_r)$ positive and decreasing in magnitude, and all other entries zero:

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}, \quad \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$$. 

Corollary 1 (Compact-form SVD)

- Any matrix $A \in \mathbb{R}^{m \times n}$ can be expressed as
  
  $$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

  where $r = \text{rank}(A)$, $U = [u_1 \cdots u_r]$ is such that $U^T U = I_r$, $V = [v_1 \cdots v_r]$ is such that $V^T V = I_r$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

- The positive numbers $\sigma_i$ are called the singular values of $A$, vectors $u_i$ are called the left singular vectors of $A$, and $v_i$ the right singular vectors. These quantities satisfy
  
  $$AV_i = \sigma_i u_i, \quad A^T u_i = \sigma_i v_i, \quad i = 1, \ldots, r.$$ 

- Moreover, $\sigma_i^2 = \lambda_i(AA^T) = \lambda_i(A^T A)$, $i = 1, \ldots, r$, and $u_i, v_i$ are the eigenvectors of $A^T A$ and of $AA^T$, respectively.


decayed and transcribed content

Notes

Interpretation

The singular value decomposition theorem allows to write any matrix can be written as a sum of dyads:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

where

- Vectors $u_i, v_i$ are normalized, with $\sigma_i > 0$ providing the "strength" of the corresponding dyad;
- The vectors $u_i, i = 1, \ldots, r$ (resp. $v_i, i = 1, \ldots, r$) are mutually orthogonal.

Notes

Matrix properties via SVD

Rank, nullspace and range

- The rank $r$ of $A$ is the cardinality of the nonzero singular values, that is the number of nonzero entries on the diagonal of $\Sigma$.
- Since $r = \text{rank}(A)$, by the fundamental theorem of linear algebra the dimension of the nullspace of $A$ is $\dim(N(A)) = n - r$. An orthonormal basis spanning $N(A)$ is given by the last $n-r$ columns of $V$, i.e.
  
  $$N(A) = \mathbb{R}(V_{n-r}) = \mathbb{R}([v_{r+1} \cdots v_n]).$$

- Similarly, an orthonormal basis spanning the range of $A$ is given by the first $r$ columns of $U$, i.e.
  
  $$\mathbb{R}(A) = \mathbb{R}(U_r) = \mathbb{R}([u_1 \cdots u_r]).$$

Notes

Matrix properties via SVD

Matrix norms

- The squared Frobenius matrix norm of a matrix $A \in \mathbb{R}^{m \times n}$ can be defined as
  
  $$\|A\|_F^2 = \text{trace}(A^T A) = \sum_{i=1}^{r} \lambda_i(A^T A) = \sum_{i=1}^{r} \sigma_i^2,$$

  where $\sigma_i$ are the singular values of $A$. Hence the squared Frobenius norm is nothing but the sum of the squares of the singular values.

- The squared spectral matrix norm $\|A\|_2^2$ is equal to the maximum eigenvalue of $A^T A$, therefore
  
  $$\|A\|_2^2 = \sigma_1^2,$$

  i.e., the spectral norm of $A$ coincides with the maximum singular value of $A$.

- The so-called nuclear norm of a matrix $A$ is defined in terms of its singular values:
  
  $$\|A\|_\ast = \sum_{i=1}^{r} \sigma_i, \quad r = \text{rank}(A)$$

  The nuclear norm appears in several problems related to low-rank matrix completion or rank minimization problems.
Matrix properties via SVD

Condition number
- The condition number of an invertible matrix $A \in \mathbb{R}^{n \times n}$ is defined as the ratio between the largest and the smallest singular value:
  \[ \kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}} = ||A|| \cdot ||A^{-1}||. \]
- This number provides a quantitative measure of how close $A$ is to being singular (the larger $\kappa(A)$ is, the more close to singular $A$ is).
- The condition number also provides a measure of the sensitivity of the solution of a system of linear equations to changes in the equation coefficients.

Matrix properties via SVD

Matrix pseudo-inverses
- Given $A \in \mathbb{R}^{m \times n}$, a pseudoinverse is a matrix $A^+$ that satisfies
  \[ AA^+A = A, \quad A^+AA^+ = A^+ \]
  \[ (AA^+)^T = AA^+, \quad (A^+A)^T = A^+A. \]
- A specific pseudoinverse is the so-called Moore-Penrose pseudoinverse:
  \[ A^+ = V\tilde{\Sigma}U^* \in \mathbb{R}^{m \times n} \]
  where
  \[ \tilde{\Sigma} = \begin{bmatrix} \Sigma^{-1} & 0_{m-r} \\ 0_{r-m} & 0_{r-m} \end{bmatrix}, \quad \Sigma^{-1} = \text{diag}(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}) > 0. \]
- Due to the zero blocks in $\tilde{\Sigma}$, $A^+$ can be written compactly as
  \[ A^+ = V\Sigma^{-1}U^*. \]

Moore-Penrose pseudo-inverse: special cases
- If $A$ is square and nonsingular, then $A^+ = A^{-1}$.
- If $A \in \mathbb{R}^{m \times n}$ is full column rank, that is $r = n \leq m$, then $A^+A = VV^* = I_n,$
  that is, $A^+$ is a left inverse of $A$, and it has the explicit expression
  \[ A^+ = (A^+A)^{-1}A^+. \]
- If $A \in \mathbb{R}^{m \times n}$ is full row rank, that is $r = m \leq n$, then $AA^+ = UU^* = I_n,$
  that is, $A^+$ is a right inverse of $A$, and it has the explicit expression
  \[ A^+ = A^+(AA^+)^{-1}. \]

Matrix properties via SVD

Projectors
- Matrix $P_{R(A)} = A^+A = U\Sigma^T$ is an orthogonal projector onto $R(A)$. This means that, for any $y \in \mathbb{R}^n$, the solution of problem
  \[ \min_{z \in R(A)} ||z - y|| \]
  is given by $z^* = P_{R(A)}y$.
- Similarly, matrix $P_{N(A)^\perp} = (I_n - AA^+)$ is an orthogonal projector onto $N(A)^\perp = \mathbb{R}(A^\perp)$.
- Matrix $P_{N(A^\perp)} = I_n - A^+A$ is an orthogonal projector onto $N(A)$.
- Matrix $P_{R(A^\perp)} = A^+A$ is an orthogonal projector onto $\mathbb{R}(A)^\perp = \mathbb{R}(A^\perp)$. 

Notes
Low-rank matrix approximation

Let \( A \in \mathbb{R}^{m \times n} \) be a given matrix, with \( \text{rank}(A) = r > 0 \). We consider the problem of approximating \( A \) with a matrix of lower rank. In particular, we consider the following rank-constrained approximation problem

\[
\min_{A_k \in \mathbb{R}^{m \times n}} \| A - A_k \|_F^2 \\
\text{s.t.: } \text{rank}(A_k) = k,
\]

where \( 1 \leq k \leq r \) is given.

Let \( A = U \tilde{\Sigma} V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \) be an SVD of \( A \). An optimal solution of the above problem is simply obtained by truncating the previous summation to the \( k \)-th term, that is

\[
A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.
\]

The ratio

\[
\eta_k = \frac{\| A - A_k \|_F^2}{\| A \|_F^2} = \frac{\sigma_{k+1}^2 + \cdots + \sigma_r^2}{\sigma_1^2 + \cdots + \sigma_r^2}
\]

indicates what fraction of the total variance (Frobenius norm) in \( A \) is explained by the rank \( k \) approximation of \( A \).

A plot of \( \eta_k \) as a function of \( k \) may give useful indications on a good rank level \( k \) at which to approximate \( A \).

\( \eta_k \) is related to the relative norm approximation error

\[
e_k = \frac{\| A - A_k \|_F^2}{\| A \|_F^2} = \frac{\sigma_{k+1}^2 + \cdots + \sigma_r^2}{\sigma_1^2 + \cdots + \sigma_r^2} = 1 - \eta_k.
\]

Minimum “distance” to rank deficiency

Suppose \( A \in \mathbb{R}^{m \times n} \), \( m \geq n \) is full rank, i.e., \( \text{rank}(A) = n \). We ask what is a minimal perturbation \( \delta A \) of \( A \) that makes \( A + \delta A \) rank deficient. The Frobenius norm (or the spectral norm) of the minimal perturbation \( \delta A \) measures the “distance” of \( A \) from rank deficiency.

Formally, we need to solve

\[
\min_{\delta A \in \mathbb{R}^{m \times n}} \| \delta A \|_F^2 \\
\text{s.t.: } \text{rank}(A + \delta A) = n - 1.
\]

This problem is equivalent to rank approximation, for \( \delta A = A_k - A \). The optimal solution is thus readily obtained as

\[
\delta A^* = A_k - A,
\]

where \( A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T \). Therefore, we have

\[
\delta A^* = -\sigma_{k+1} u_{k+1} v_{k+1}^T.
\]

The minimal perturbation that leads to rank deficiency is a rank-one matrix. The distance to rank deficiency is

\[
\| \delta A^* \|_F = \| \delta A^* \|_2 = \sigma_{k+1}.
\]

Example: Image compression

A 266 \times 400 matrix \( A \) of integers corresponding to the gray levels of the pixels in an image.

Compute the SVD of matrix \( A \), and plot the ratio \( \eta_k \), for \( k \) from 1 to 266

\( k = 9 \) already captures 96% of the image variance.
Example: Image compression

![Image of images with different approximations]

**Figure:** Rank $k$ approximations of the original image, for $k = 9$ (top left), $k = 23$ (top right), $k = 49$ (bottom left), and $k = 154$ (bottom right).

Principal component analysis (PCA)

- Principal component analysis (PCA) is a technique of unsupervised learning, widely used to “discover” the most important, or informative, directions in a data set, that is the directions along which the data varies the most.
- In the data cloud below it is apparent that there exist a direction (at about 45 degrees from the horizontal axis) along which almost all the variation of the data is contained. In contrast, the direction at about 135 degrees contains very little variation of the data.

The important direction was easy to spot in this two-dimensional example. However, graphical intuition does not help when analyzing data in dimension $n > 3$, which is where Principal Components Analysis (PCA) comes in handy.

Principal component analysis (PCA)

- $x_i \in \mathbb{R}^n$, $i = 1, \ldots, m$: data points; $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i$: the barycenter of the data points; $\tilde{X} \in \mathbb{R}^{m \times n}$ matrix containing the centered data points: $\tilde{x}_i = x_i - \bar{x}$, $\bar{x} = x_i - \bar{x}$, $i = 1, \ldots, m$.
- We look for a normalized direction in data space, $z \in \mathbb{R}^n$, $|z| = 1$, such that the variance of the projections of the centered data points on the line determined by $z$ is maximal.
- The components of the centered data along direction $z$ are given by $\alpha_i = \tilde{x}_i^T z$, $i = 1, \ldots, m$.
  - ($\alpha_i, z$ are the projections of $\tilde{x}_i$ along the span of $z$).
  - The mean-square variation of the data along direction $z$ is thus given by
    $$\frac{1}{m} \sum_{i=1}^{m} \alpha_i^2 = \sum_{i=1}^{m} \tilde{x}_i^T \tilde{x}_i z = z^T \tilde{X}^T \tilde{X} z.$$

Successive principal axes can be found by “removing” the first principal components, and applying the same approach again on the “deflated” data matrix.

Principal component analysis (PCA)

- The direction $z$ along which the data has the largest variation can thus be found as the solution to the following optimization problem:
  $$\max_{z \in \mathbb{R}^n} z^T (\tilde{X} \tilde{X}^T) z \quad \text{s.t.} \quad |z| = 1.$$
- Let us now solve the previous problem via the SVD of $\tilde{X}$: let
  $$\tilde{X} = U \Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T.$$
- Then, $H = \tilde{X} \tilde{X}^T = U \Sigma^2 U^T$.
- From the variational representation, we have that the optimal solution of this problem is given by the column $u_1$ of $U$ corresponding to the largest eigenvalue of $H$, which is $\sigma_1^2$.
- The direction of largest data variation is thus readily found as $z = u_1$, and the mean-square variation along this direction is proportional to $\sigma_1^2$.
- Successive principal axes can be found by “removing” the first principal components, and applying the same approach again on the “deflated” data matrix.
Example: PCA of market data

Consider data consisting in the returns of six financial indices: (1) the MSCI US index, (2) the MSCI EUR index, (3) the MSCI JAP index, (4) the MSCI PACIFIC index, the (5) MSCI BOT liquidity index, and the (6) MSCI WORLD index.

We used monthly return data, from Feb. 26, 1993 to Feb. 28, 2007, for a total of 169 data points.

The data matrix $X$ has thus $m = 169$ data points in dimension $n = 6$.

Centering the data, and performing the SVD on the centered data matrix $\tilde{X}$, we obtain the principal axes $u_i$ and the corresponding singular values:

$$ U = \begin{bmatrix} 0.6405 & 0.2087 & 0.3333 & 0.0730 & 0.0511 & 0.0049 \\ 0.0559 & 0.6648 & 0.2498 & 0.0579 & 0.0804 & 0.0056 \\ 0.1085 & 0.3862 & 0.0402 & 0.0323 & 0.0238 & 0.0211 \\ 0.0593 & 0.5344 & 0.1272 & 0.0395 & 0.0204 & 0.0036 \\ 0.0203 & 0.1012 & 0.0519 & 0.0160 & 0.0094 & 0.0009 \\ 0.1168 & 0.3959 & 0.0244 & 0.0180 & 0.0022 & 0.0010 \end{bmatrix} $$

Computing the ratios $\eta_k$, we have

$$ \eta \times 100 = [67.77 \ 84.58 \ 96.21 \ 99.92 \ 99.99 \ 100] $$

We deduce, for instance, that over 96% of the variability in the returns of these six assets can be explained in terms of only three implicit “factors”.

In statistical terms, this means that each realization of the return vector $x \in \mathbb{R}^6$ can be expressed (up to a 96% “approximation”) as

$$ x = \bar{x} + U_3 z, $$

where $z$ is a zero-mean vector of random factors, and $U_3 = [u_1, u_2, u_3]$ is the factor loading matrix, composed of the first three principal directions of the data.

Example: PCA of voting data

Figure: Projection of US Senate voting data on random direction (left panel) and direction of maximal variance (right panel). The latter reveals party structure (party affiliations added after the fact). Note also the much higher range of values it provides.