Dyads

rank 1 matrix

A matrix $A \in \mathbb{R}^{m,n}$ 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

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

Interpretation:

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

Dyads

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.

Figure: Row vector representation of 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 intepret data matrices as sums of "simpler" matrices (dyads).

The singular value decomposition (SVD)

SVD theorem

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 $A \in \mathbb{R}^{m,n}$.

Theorem 1 (SVD decomposition)

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

$$A = U\tilde{\Sigma}V^\top$$

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

$$\tilde{\Sigma} = \left[\begin{array}{cc} \Sigma & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{array} \right], \quad \Sigma = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_r) \succ 0.$$

Compact-form SVD

Corollary 1 (Compact-form SVD)

- *Any matrix* $A \in \mathbb{R}^{m,n}$ *can be expressed as*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top = U_r \Sigma V_r^\top$$

where $r = \operatorname{rank} A$, $U_r = [u_1 \ \cdots \ u_r]$ *is such that* $U_r^\top U_r = I_r$, $V_r = [v_1 \ \cdots \ v_r]$ *is such that* $V_r^\top V_r = I_r$, *and* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

- *The positive numbers* σ_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*

$$Av_i = \sigma_i u_i, \quad u_i^\top A = \sigma_i v_i, \qquad i = 1, \ldots, r.$$

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

Interpretation

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

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

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.

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 $\tilde{\Sigma}$.
- Since $r = \mathrm{rank}\, A$, by the fundamental theorem of linear algebra the dimension of the nullspace of A is $\dim \mathcal{N}(A) = n - r$. An orthonormal basis spanning $\mathcal{N}(A)$ is given by the last $n - r$ columns of V, i.e.

$$\mathcal{N}(A) = \mathcal{R}(V_{nr}), \quad V_{nr} \doteq [v_{r+1} \ \cdots \ v_n].$$

- Similarly, an orthonormal basis spanning the range of A is given by the first r columns of U, i.e.

$$\mathcal{R}(A) = \mathcal{R}(U_r), \quad U_r \doteq [u_1 \ \cdots \ u_r].$$

Matrix properties via SVD

Matrix norms

- The squared Frobenius matrix norm of a matrix $A \in \mathbb{R}^{m,n}$ can be defined as

$$\|A\|_F^2 = \operatorname{trace} A^\top A = \sum_{i=1}^{n} \lambda_i(A^\top A) = \sum_{i=1}^{n} \sigma_i^2,$$

  where σ_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 $A^\top A$, 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:

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i, \quad r = \operatorname{rank} A.$$

  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 $A \in \mathbb{R}^{n,n}$ is defined as the ratio between the largest and the smallest singular value:

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} = \|A\|_2 \cdot \|A^{-1}\|_2.$$

- 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 $A \in \mathbb{R}^{m,n}$, a *pseudoiverse* is a matrix $A^\dagger$ that satisfies

$$\begin{array}{rcl} AA^\dagger A & = & A \\ A^\dagger AA^\dagger & = & A^\dagger \\ (AA^\dagger)^\top & = & AA^\dagger \\ (A^\dagger A)^\top & = & A^\dagger A. \end{array}$$

- A specific pseudoinverse is the so-called Moore-Penrose pseudoinverse:

$$A^\dagger = V\tilde{\Sigma}^\dagger U^\top \in \mathbb{R}^{n,m}$$

where

$$\tilde{\Sigma}^\dagger = \left[\begin{array}{cc} \Sigma^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{array} \right], \quad \Sigma^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}\right) \succ 0.$$

- Due to the zero blocks in $\tilde{\Sigma}$, $A^\dagger$ can be written compactly as

$$A^\dagger = V_r \Sigma^{-1} U_r^\top.$$

Matrix properties via SVD

Moore-Penrose pseudo-inverse: special cases

- If A is square and nonsingular, then $A^\dagger = A^{-1}$.
- If $A \in \mathbb{R}^{m,n}$ is full column rank, that is $r = n \leq m$, then

$$A^\dagger A = V_r V_r^\top = VV^\top = I_n,$$

  that is, $A^\dagger$ is a *left inverse* of A, and it has the explicit expression

$$A^\dagger = (A^\top A)^{-1} A^\top.$$

- If $A \in \mathbb{R}^{m,n}$ is full row rank, that is $r = m \leq n$, then

$$AA^\dagger = U_r U_r^\top = UU^\top = I_m,$$

  that is, $A^\dagger$ is a *right inverse* of A, and it has the explicit expression

$$A^\dagger = A^\top (AA^\top)^{-1}.$$

Matrix properties via SVD

Projectors

- Matrix $P_{\mathcal{R}(A)} \doteq AA^\dagger = U_r U_r^\top$ is an orthogonal projector onto $\mathcal{R}(A)$. This means that, for any $y \in \mathbb{R}^n$, the solution of problem

$$\min_{z \in \mathcal{R}(A)} \|z - y\|_2$$

  is given by $z^* = P_{\mathcal{R}(A)} y$.
- Similarly, matrix $P_{\mathcal{N}(A^\top)} \doteq (I_m - AA^\dagger)$ is an orthogonal projector onto $\mathcal{R}(A)^\perp = \mathcal{N}(A^\top)$.
- Matrix $P_{\mathcal{N}(A)} \doteq I_n - A^\dagger A$ is an orthogonal projector onto $\mathcal{N}(A)$.
- Matrix $P_{\mathcal{N}(A)^\perp} \doteq A^\dagger A$ is an orthogonal projector onto $\mathcal{N}(A)^\perp = \mathcal{R}(A^\top)$.

Low-rank matrix approximation

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

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

  where $1 \leq k \leq r$ is given.

$\|A\|_F^2 = \operatorname{trace}(A^T A)$

$= \sum \sigma^2_{A^TA} = \sum \lambda_{A^TA}$

- Let

$$A = U\tilde{\Sigma}V^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

  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

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top.$$

Low-rank matrix approximation

- The ratio

$$\eta_k = \frac{\|A_k\|_F^2}{\|A\|_F^2} = \frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \cdots + \sigma_r^2}$$

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

- A plot of η_k as a function of k may give useful indications on a good rank level k at which to approximate A.

- η_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 $A \in \mathbb{R}^{m,n}$, $m \geq n$ is full rank, i.e., $\mathrm{rank}(A) = n$. We ask what is a minimal perturbation δA of A that makes $A + \delta A$ rank deficient. The Frobenius norm (or the spectral norm) of the minimal perturbation δA measures the “distance” of A from rank deficiency.
- Formally, we need to solve

$$\begin{aligned} \min_{\delta A \in \mathbb{R}^{m,n}} \quad & \|\delta A\|_F^2 \\ \text{s.t.:} \quad & \mathrm{rank}(A + \delta A) = n - 1. \end{aligned}$$

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

$$\delta A^* = A_k - A,$$

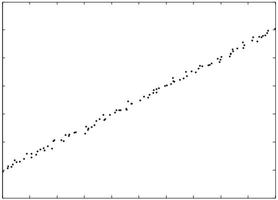
where $A_k = \sum_{i=1}^{n-1} \sigma_i u_i v_i^\top$. Therefore, we have

$$\delta A^* = -\sigma_n u_n v_n^\top.$$

- The minimal perturbation that leads to rank deficiency is a rank-one matrix. The distance to rank deficiency is $\|\delta A^*\|_F = \|\delta A^*\|_2 = \sigma_n$.

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; $\tilde{X}$ $n \times m$ matrix containing the centered data points:

$$\tilde{X} = [\tilde{x}_1 \ \cdots \ \tilde{x}_m], \quad \tilde{x}_i \doteq x_i - \bar{x}, \ i = 1, \ldots, m,$$

- We look for a a normalized direction in data space, $z \in \mathbb{R}^n$, $\|z\|_2 = 1$, such that the 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

$$\alpha_i = \tilde{x}_i^\top z, \quad i = 1, \ldots, m.$$

($\alpha_i z$ 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

$$\frac{1}{m}\sum_{i=1}^m \alpha_i^2 = \sum_{i=1}^m z^\top \tilde{x}_i \tilde{x}_i^\top z = z^\top \tilde{X}\tilde{X}^\top z.$$

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:

$$\max_{z \in \mathbb{R}^n} \quad z^\top (\tilde{X}\tilde{X}^\top) z \quad \text{s.t.: } \|z\|_2 = 1.$$

- Let us now solve the previous problem via the SVD of $\tilde{X}$: let

$$\tilde{X} = U_r \Sigma V_r^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$

- Then, $H \doteq \tilde{X}\tilde{X}^\top = U_r \Sigma^2 U_r^\top$.
- From the variational representation, we have that the optimal solution of this problem is given by the column u_1 of U_r corresponding to the largest eigenvalue of H, which is σ_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 σ_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.