Foundations of Data Analysis (MA4800)

Massimo Fornasier

Fakultät für Mathematik
Technische Universität München
massimo.fornasier@ma.tum.de
http://www-m15.ma.tum.de/

Slides of Lecture
May 9 2017

Preliminaries on Linear Algebra (LA)

We give for granted familiarity with the basics of LA taught in standard courses, in particular,

- vector spaces, spans and linear combinations, linear bases, linear maps and matrices, eigenvalues, complex numbers, scalar products, theory of symmetric matrices.

For more details we refer to the lecture notes (in German)

G. Kemper, *Lineare Algebra fuer Informatik*, TUM, 2017.

of the course taught for Computer Scientsts at TUM, or any other standard international text of LA.

Matrix Notations

The index sets $I, J, L, \ldots$ are assumed to be finite.

As soon as the complex conjugate values appear[1], the scalar field is restricted to $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

The set $\mathbb{K}^{I \times J}$ is the vector space of matrices $A \in \mathbb{K}^{I \times J}$, whose entries are denoted by A_{ij}, for $i \in I$ and $j \in J$. Vice versa, numbers $a_{ij} \in \mathbb{K}$ may be used to define $A := (a_{ij})_{i \in I, j \in J} \in \mathbb{K}^{I \times J}$.

If $A \in \mathbb{K}^{I \times J}$, the transposed matrix $A^T \in \mathbb{K}^{J \times I}$, and $A^T_{ji} := A_{ij}$. A matrix $A \in \mathbb{K}^{I \times I}$ is symmetric if $A^T = A$. The Hermitian transposed matrix $A^H \in \mathbb{K}^{J \times I}$ coincides with $\overline{A^T}$. If $\mathbb{K} = \mathbb{R}$ then clearly $A^H = A^T$. A Hermitian matrix satisfies $A^H = A$.

Often, we will need to consider the rows $A^{(i)}$ and the columns $A_{(j)}$ of a matrix, defined respectively as vectors $A^{(i)} = (a_{ij})_{j \in J}$ and $A_{(j)} = (a_{ij})_{i \in I} = (A^T)^{(j)}$.

---

[1] In the case $\mathbb{K} = \mathbb{R}$, then $\alpha = \overline{\alpha}$, for all $\alpha \in \mathbb{K}$.

Matrix Notations

We assume that the standard matrix-vector and matrix-matrix multiplications are done in the usual way: $(Ax)_i = \sum_{j\in J} a_{ij}x_j$ or $(AB)_{i\ell} = \sum_{j\in J} A_{ij}B_{j\ell}$ as usual, for $x \in \mathbb{K}^J$, $A \in \mathbb{K}^{I\times J}$ and $B \in \mathbb{K}^{J\times L}$.

The Kronecker symbol is defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \in I, \\ 0, & \text{otherwise.} \end{cases}$$

The unit vector $e^{(i)} \in \mathbb{K}^I$ is defined by

$$e^{(i)} = (\delta_{ij})_{j\in I}$$

The symbol $I = (\delta_{ij})_{j\in I, j\in I}$ is used for the identity matrix. Since matrices and index sets do not appear in the same place, the simultaneous use of the symbol I do not create confusion (example: $I \in \mathbb{K}^{I\times I}$).

Matrix Notations

The range of matrix $A \in \mathbb{K}^{I \times J}$ is

$$\text{range}(A) = \{Ax : x \in \mathbb{K}^J\} = \text{span}\{A_{(i)}, i \in I\}.$$

Hence, the range of a matrix is a vector space spanned by its columns.

The Euclidean scalar product in $\mathbb{K}^I$ is given by

$$\langle x, y \rangle = y^H x = \sum_{i \in I} x_i \overline{y}_i, \tag{1}$$

where for $\mathbb{K} = \mathbb{R}$ the conjugate sign can be ignored .

It often useful and very important to notice that the matrix-vector Ax and matrix-matrix AB multiplications can also be expressed in terms of scalar products of the rows of A with x and of the rows of A with the colums of B, i.e., according to our terminology of

$$(Ax)_i = \langle A^{(i)}, \overline{x} \rangle, \quad (AB)_{i\ell} = \langle A^{(i)}, \overline{B_{(\ell)}} \rangle. \tag{2}$$

Matrix Notations

Two vectors $x, y \in \mathbb{K}^I$ are orthogonal (and we may write $x \perp y$) if $\langle x, y \rangle = 0$.

We often consider sets which are mutually orthogonal. In this case for $X, Y \subset \mathbb{K}^I$ we say that X is orthogonal to Y and we write $X \perp Y$ if $\langle x, y \rangle = 0$, for all $x \in X$ and $y \in Y$.

When X and Y are linear subspaces their orthogonality can be simply checked by showing that they possess bases which are mutually orthogonal. This is a very important principle we will use often.

A family of vectors $X = \{x_\nu\}_{\nu \in F} \subset \mathbb{K}^I$ is orthogonal if the vectors x_ν are pairwise orthogonal, i.e., $\langle x_\nu, x_{\nu'} \rangle = 0$ for $\nu \neq \nu'$. The family is additionally called orthonormal if $\langle x_\nu, x_\nu \rangle = 1$ for all $\nu \in F$.

Matrix Notations

A matrix $A \in \mathbb{K}^{I \times J}$ is called orthogonal, if the columns of A are orthonormal, equivalently if

$$A^H A = I \in \mathbb{K}^{J \times J}.$$

An orthogoal square matrix $A \in \mathbb{K}^{I \times I}$ is called unitary. Differently from just orthogonal matrices, unitary matrices satisfy

$$A^H A = A A^H = I,$$

i.e, $A^H = A^{-1}$ is the invese matrix of A.

Assume that the index sets satisfy either $I \subset J$ or $J \subset I$. Then a (rectangular) matrix $A \in \mathbb{K}^{I \times J}$ is diagonal if $A_{ij} = 0$ for all $i \neq j$.

Matrix Rank

Proposition

Let $A \in \mathbb{K}^{I \times J}$. The following statements are equivalent and may be all used as a definition of matrix rank $r = \operatorname{rank}(A)$.

- *(a) $r = \dim \operatorname{range}(A)$;*
- *(b) $r = \dim \operatorname{range}(A^H)$;*
- *(c) r is the maximal number of linearly independent rows of A;*
- *(d) r is the maximal number of linearly independent columns of A;*
- *(e) r is minimal with the property*

$$A = \sum_{i=1}^{r} a_i b_i^H, \text{ where } a_i \in \mathbb{K}^I, b \in \mathbb{K}^J;$$

- *(f) r is maximal with the property that there exists an invertible $r \times r$ submatrix of A;*
- *(g) r is the number of positive singular values (soon!).*

Matrix Rank

The rank is bounded by the maximal rank, which, for matrices, is always given by $r_{max} = \min\{\#I, \#J\}$, and this bound is attained by full-rank matrices.

As linear independence may depend on the field $\mathbb{K}$ one may question whether the rank of a real-valued matrix depends on considering it in $\mathbb{R}$ or in $\mathbb{C}$. For matrices it tunrs out that it does not matter: the rank is the same.

We will often work with matrices of bounded rank $r \leq k$. We denote accordingly with $\mathcal{R}_k = \{A \in \mathbb{K}^{I \times J} : \text{range}(A) \leq k\}$ such a set. Notice that this set is not a vector space (exercise!).

Norms

In the following we consider an abstract vector space V over the field $\mathbb{K}$. As a typical example we may keep in mind the Euclidean plane $V = \mathbb{R}^2$ endowed with the classical Euclidean norm.

We recall below the axioms of an abstract norm on V: a norm is a map $\|\cdot\| : V \to [0, \infty)$ with the following properties

- $\|v\| = 0$ if and only if $v = 0$;
- $\|\lambda v\| = |\lambda| \|v\|$ for all $v \in V$ and $\lambda \in \mathbb{K}$;
- $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in V$ (triangle inequality).

A norm is always continuous as a consequence of the (inverse) triangle inequalty:

$$|\|v\| - \|w\|| \leq \|v - w\|, \text{ for all } v, w \in V.$$

A vector space V endowed with a norm $\|\cdot\|$, and we write the pair $(V, \|\cdot\|)$ to indicate it, is called a normed vector space. As mentioned above a typical example is the Euclidean plane.

Scalar products and (pre)-Hilbert spaces

A normed vector space $(V, \|\cdot\|)$ is a pre-Hilbert space if its norm is defined by

$$\|v\| = \sqrt{\langle v, v\rangle}, \quad v \in V,$$

where $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{K}$ is a scalar product on V, i.e., it fulfills the properties

- $\langle v, v\rangle > 0$ for $v \neq 0$;
- $\langle v, w\rangle = \overline{\langle w, v\rangle}$, for $v, w \in V$;
- $\langle u + \lambda v, w\rangle = \langle u, w\rangle + \lambda\langle v, w\rangle$ for $u, v, w \in V$ and $\lambda \in \mathbb{K}$;
- $\langle w, u + \lambda v\rangle = \langle w, u\rangle + \overline{\lambda}\langle w, v\rangle$ for $u, v, w \in V$ and $\lambda \in \mathbb{K}$.

The triangle inequality for the norm follows from the Schwarz inequality

$$|\langle v, w\rangle| \leq \|v\|\|w\|, \quad v, w \in V.$$

Scalar products and (pre)-Hilbert spaces

We describe the pre-Hilbert space also by the pair $(V, \langle \cdot, \cdot \rangle)$. A typical example of scalar product over $\mathbb{K}^I$ is the one we introduced in (1), which generates the Euclidean norm on $\mathbb{K}^I$:

$$\|v\|_2 = \sqrt{\sum_{i \in I} |v_i|^2}, \quad v \in \mathbb{K}^I.$$

As before one can define orthogonality between vectors, orthogonal, and orthonormal sets of vectors.

Hilbert spaces

A Hilbert space is a pre-Hilbert space $(V, \langle \cdot, \cdot \rangle)$, whose topology induced by the associated norm $\| \cdot \|$ is complete, i.e., any Cauchy sequence in such vector spaces is convergent.

It is important to stress that every finite dimensional pre-Hilbert space $(V, \langle \cdot, \cdot \rangle)$ is actually complete, hence always a Hilbert space.

Thus, those who are not familiar with topological notions (e.g., completeness, Cauchy sequences, etc.), one should just reason in what follows according to their Euclidean geometric intuition.

Projections

Recall now that a set C in a vector space is convex if, for all $v, w \in C$ and all $t \in [0,1]$, $tv + (1 - tw) \in C$.

Given a closed convex set C in a Hilbert space $(V, \langle \cdot, \cdot \rangle)$ one defines the projection of any vector v on C as

$$P_C(v) = \arg\ \min_{w \in C} \|v - w\|.$$

This definition is well-posed, as the projection is actually unique, and an equivalent defition is given by fulfilling the following inequality

$$\langle z - P_C(v), v - P_C(v) \rangle \leq 0,$$

for all $z \in C$. This is left as an exercise.

Projections onto subspaces: Pythagoras-Fourier Theorem

Of extreme importance for us are the orthogonal projections onto subspaces.

In case $C = W \subset V$ is actually a closed linear subspace of V, then the projection onto W can be readily computed as soon as one disposes of an orthonormal basis for W. Let $\{w_\nu\}_{\nu \in F}$ be a (countable) orthonormal basis for W then

$$P_W(v) = \sum_{\nu \in F} \langle v, w_\nu \rangle w_\nu, \text{ for all } v \in V. \tag{3}$$

Moreover, it holds the Pythagoras-Fourier Theorem:

$$\|P_W(v)\|^2 = \sum_{\nu \in F} |\langle v, w_\nu \rangle|^2, \text{ for all } v \in V.$$

Pythagoras-Fourier Theorem and othonormal expansions

We will use very much this characterization of the orthogonal projections and the Pythagoras-Fourier Theorem, especially for the case where $W = V$.

In this case, obviosly, $P_V = I$ and, we have the orthonormal expansion of any vector $v \in V$,

$$v = \sum_{\nu \in F} \langle v, w_\nu \rangle w_\nu,$$

and the norm equivalence

$$\|v\|^2 = \sum_{\nu \in F} |\langle v, w_\nu \rangle|^2.$$

Trace of a matrix

The effort of defining a abstract scalar products and norms allows us now to introduce several norms for matrices.

First of all we need to introduce the concept of trace, which is the map $\mathrm{tr} : \mathbb{K}^{I\times I} \to \mathbb{K}$ defined by

$$\mathrm{tr}(A) = \sum_{i\in I} A_{ii},$$

i.e., it is the sum of the diagonal elements of the matrix A.

Trace of a matrix: properties

The trace enjoyes several properties, which we collect in the following:

Proposition (Properties of the trace)

(a) $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, *for any* $A \in \mathbb{K}^{I\times J}$ *and* $B \in \mathbb{K}^{J\times I}$;

(b) $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$, *for any* A, B, C *matrices of compatible size and indexes; this property is called the circularity property of the trace;*

(c) *as a consequence of the previous property we obtain the invariance of the trace under unitary transformations, i.e.,* $\mathrm{tr}(A) = \mathrm{tr}(UAU^H)$ *for* $A \in \mathbb{K}^{I\times I}$ *and any unitary matrix* $U \in \mathbb{K}^{I\times I}$;

(d) $\mathrm{tr}(A) = \sum_{i\in I} \lambda_i$, *where* $\{\lambda_i : i \in I\}$ *is the set of eigenvalues of* A.

Matrix norms

The Frobenius norm of a matrix is essentially the Euclidean norm computed over the entries of the matrix (considered as a vector):

$$\|A\|_F := \sqrt{\sum_{i\in I, j\in J} |A_{ij}|^2}, \quad A \in \mathbb{K}^{I\times J}.$$

It is also know as Schur norm or Hilbert-Schmidt norm. This norm is generated by the scalar product (that's why we made the effort of introducing abstract scalar products!)

$$\langle A, B\rangle_F := \sum_{i\in I}\sum_{j\in J} A_{ij}\overline{B_{ij}} = \operatorname{tr}(AB^H) = \operatorname{tr}(B^H A). \tag{4}$$

In particular,

$$\|A\|_F^2 = \operatorname{tr}(AA^H) = \operatorname{tr}(A^H A) \tag{5}$$

holds.

Matrix norms

Let $\|\cdot\|_X$ and $\|\cdot\|_Y$ be vector norms on the vector spaces $X = \mathbb{K}^I$ and $Y = \mathbb{K}^J$, respectively. Then the associated matrix norm is

$$\|A\| := \|A\|_{X\to Y} := \sup_{z\neq 0} \frac{\|Az\|_Y}{\|z\|_X}, \quad A \in \mathbb{K}^{I\times J}.$$

If both $\|\cdot\|_X$ and $\|\cdot\|_Y$ coincides with the Euclidean norms $\|\cdot\|_2$ on $X = \mathbb{K}^I$ and $Y = \mathbb{K}^J$, respectively, then the associated matrix norm is called the spectral norm and it is denoted with $\|A\|_\infty$.

As the spectral norm has a central importance in many applications, it is often denoted just by $\|A\|$.

Unitary invariance and submultiplicativity

Both the matrix norms we introduced so far as are invariant with respect to unitary transformations, i.e., $\|A\|_F = \|UAV^H\|_F$ and $\|A\|_\infty = \|UAV^H\|_\infty$ for unitary matrices U, V.

Moreover, both are submitiplicative, i.e., $\|AB\|_F \leq \|A\|_\infty \|B\|_F \leq \|A\|_F \|B\|_F$ and $\|AB\| \leq \|A\|\|B\|$, for matrices A, B of compatible sizes.

Introduction of the Singular Value Decomposition

We introduce and analyze the singular value decomposition (SVD) of a matrix A, which is is the factorization of A into the product of three matrices $A = U\Sigma V^H$, where U, V are orthogonal matrices of compatible size and the matrix Σ is diagonal with positive real entries.

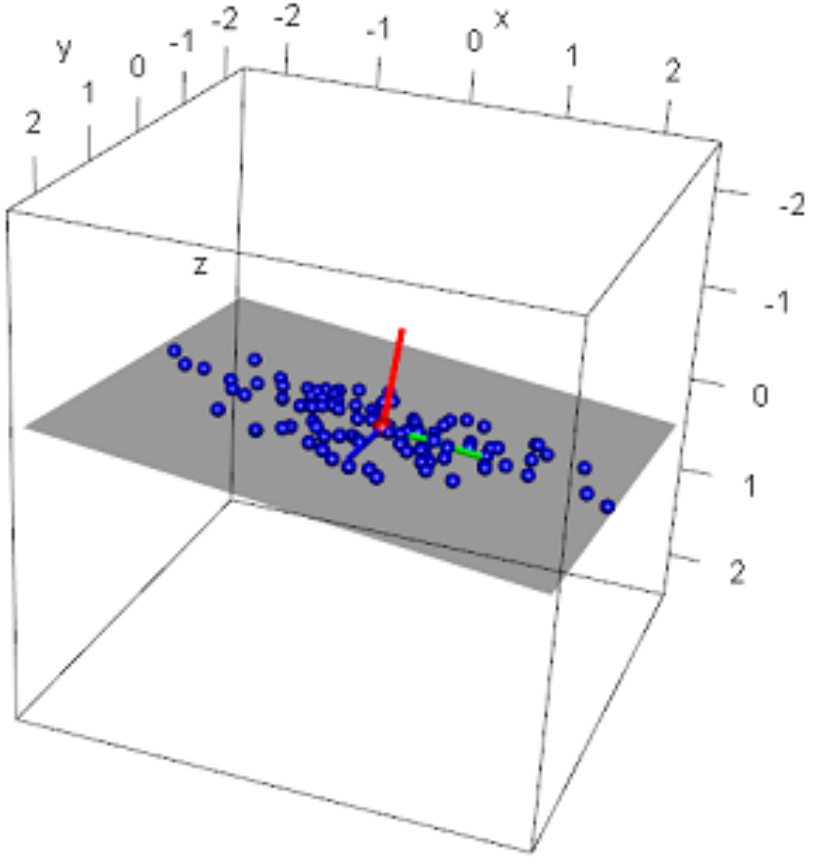
All these terms, orthogonal, diagonal matrix, have been introduced in the previous lectures.

Geometrical derivation: principal component analysis

To gain insight into the SVD, treat the $n = \#I$ rows of a matrix $A \in \mathbb{K}^{I \times J}$ as points in a d-dimensional space, where $d = \#J$, and consider the problem of finding the best k-dimensional subspace with respect to the set of points.

Here best means minimize the sum of the squares of the perpendicular distances of the points to the subspace.
An orthonormal basis for this subspace is built as fundamental directions with maximal variance of the group of high dimensional points, and are called principal components.

Geometrical derivation: principal component analysis



Pythagoras Theorem, scalar products, and orthogoal projections

We begin with a special case of the problem where the subspace is 1-dimensional, a line through the origin.

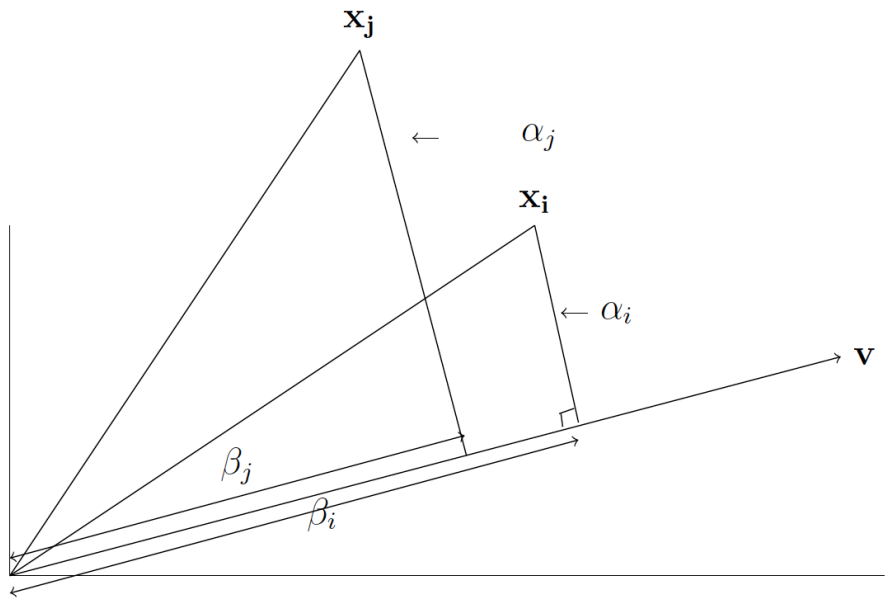
We will see later that the best-fitting k-dimensional subspace can be found by k applications of the best fitting line algorithm.

Finding the best fitting line through the origin with respect to a set of points $\{x_i := A^{(i)} \in \mathbb{K}^2 : i \in I\}$ in the Euclidean plane $\mathbb{K}^2$ means minimizing the sum of the squared distances of the points to the line. Here distance is measured perpendicular to the line. The problem is called the best least squares fit.

Pythagoras Theorem, scalar products, and orthogoal projections

In the best least squares fit, one is minimizing the distance to a subspace. Now, consider projecting orthogonally a point x_i onto a line through the origin. Then, by Pythagoras theorem

$$x_{i1}^2+x_{i2}^2+\cdots+x_{id}^2 = \text{ (length of projection) }^2+ \text{ (distance of pt. to line) }^2.$$



Pythagoras Theorem, scalar products, and orthogoal projections

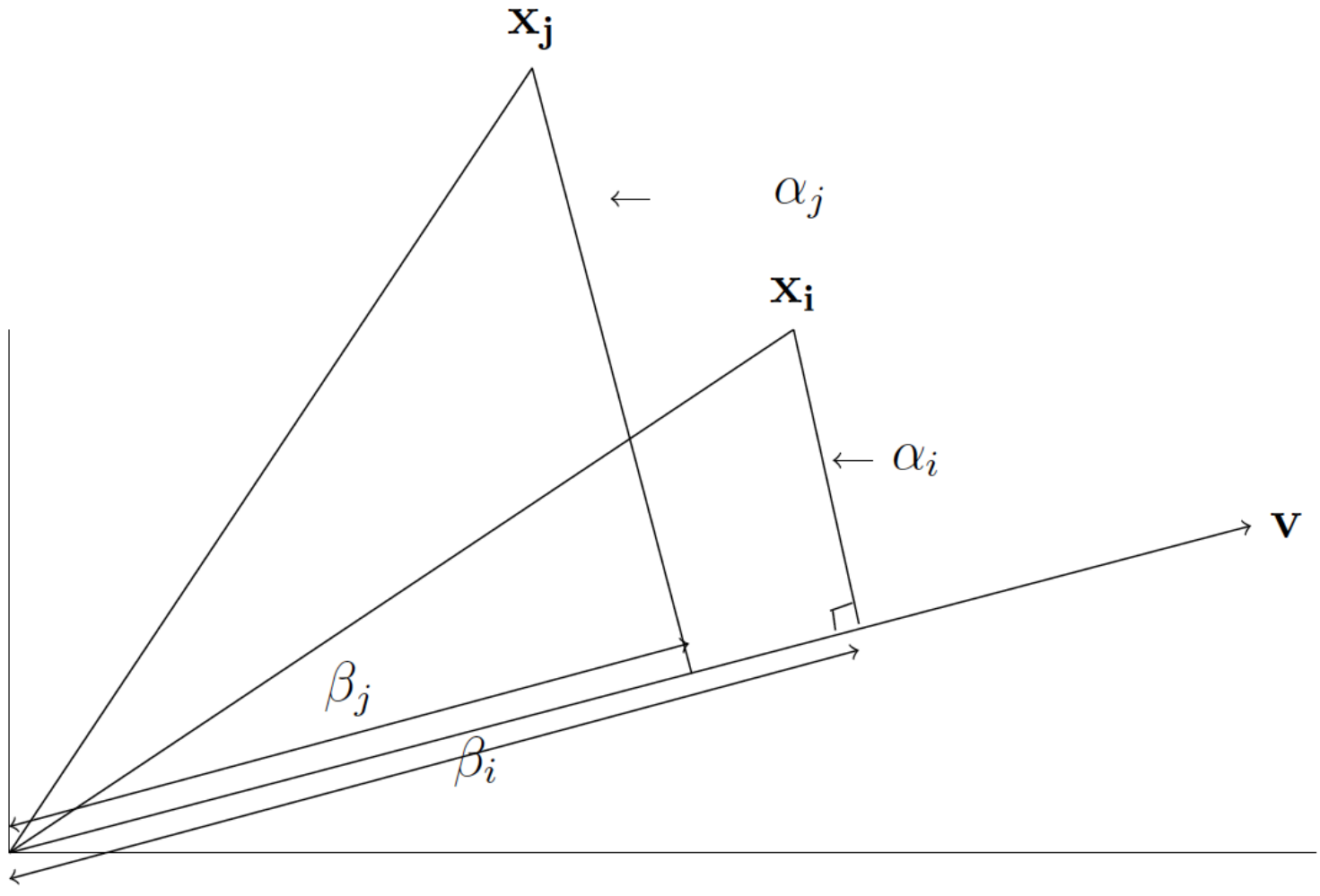
In particular, from the formula above, one has

$$(\text{distance of pt to line})^2 = x_{i1}^2 + x_{i2}^2 + \cdots + x_{id}^2 - (\text{length of projection})^2.$$

To minimize the sum of the squares of the distances to the line, one could minimize $\sum_i (x_{i1}^2 + x_{i2}^2 + \cdots + x_{id}^2)$ minus the sum of the squares of the lengths of the projections of the points to the line.

However, the first term of this difference is a constant (independent of the line), so minimizing the sum of the squares of the distances is equivalent to maximizing the sum of the squares of the lengths of the projections onto the line.

Pythagoras Theorem, scalar products, and orthogoal projections



Averiging through the points and matrix notation

For best-fit subspaces, we could maximize the sum of the squared lengths of the projections onto the subspace instead of minimizing the sum of squared distances to the subspace.

Consider the rows of A as $n = \#I$ rows of a matrix $A \in \mathbb{K}^{I \times J}$ as points in a d-dimensional space, where $d = \#J$. Consider the best-fit line through the origin.

Let v be a unit vector along this line. The length of the projection of $A^{(i)}$, the i^{th} row of A, onto v is, according to our definition of scalar product, $|\langle A^{(i)}, v \rangle|$.

From this, we see that the sum of length squared of the projections is

$$\|Av\|_2^2 = \sum_{i \in I} |\langle A^{(i)}, v \rangle|^2. \tag{6}$$

Best-fit and first singular direction

The best-fit line is the one maximizing $\|Av\|_2^2$ and hence minimizing the sum of the squared distances of the points to the line.

With this in mind, define the first singular vector, v_1 of A, which is a column vector, as the best-fit line through the origin for the n points in d-dimensional space that are the rows of A.

Thus

$$v_1 = \arg\max_{\|v\|_2=1} \|Av\|_2.$$

The value $\sigma_1(A) = \|Av_1\|_2$ is called the first singular value of A. Note that $\sigma_1(A)^2$ is the sum of the squares of the projections of the points to the line determined by v_1.

First link: linear algebra VS optimization!!!

This is the first time we use an optimization criterion to approach a data interpretation problem and to create a connection between linear algebra and optimization.

As we will discuss along this course, optimization plays indeed a huge role in data analysis and we will have to discuss it in more detail later on.

A greedy approach to best k-dimensional fit

The greedy approach to find the best-fit 2-dimensional subspace for a matrix A, takes v_1 as the first basis vector for the 2-dimenional subspace and finds the best 2-dimensional subspace containing v_1.

The fact that we are using the sum of squared distances will again help, thanks to Pythagoras Theorem: For every 2-dimensional subspace containing v_1, the sum of squared lengths of the projections onto the subspace equals the sum of squared projections onto v_1 plus the sum of squared projections along a vector perpendicular to v_1 in the subspace.

Thus, instead of looking for the best 2-dimensional subspace containing v_1, look for a unit vector, call it v_2, perpendicular to v_1 that maximizes $\|Av\|_2$ among all such unit vectors.

Using the same greedy strategy to find the best three and higher dimensional subspaces, defines $v_3, v_4, \ldots$ in a similar manner.

Formal definitions

More formally, the second singular vector, v_2 is defined by the best-fit line perpendicular to v_1:

$$v_2 = \arg\max_{\|v\|_2=1, \langle v_1, v\rangle = 0} \|Av\|_2.$$

The value $\sigma_2(A) = \|Av_2\|_2$ is called the second singular value of A.

The k^{th} singular vector v_k is defined similarly by

$$v_k = \arg\max_{\|v\|_2=1, \langle v_1, v\rangle = 0, \ldots, \langle v_{k-1}, v\rangle = 0} \|Av\|_2. \tag{7}$$

and so on.

The process stops when we have found $v_1, \ldots, v_r$ as singular vectors and

$$0 = \arg\max_{\|v\|_2=1, \langle v_1, v\rangle = 0, \ldots, \langle v_r, v\rangle = 0} \|Av\|_2.$$

Optimality of the greedy algorithm

If instead of finding v_1 that maximized $\|Av\|_2$ and then the best-fit 2-dimensional subspace containing v_1, we had found the best-fit 2-dimensional subspace, we might have done better.

Surprisingly enough, this is actually not the case. We now give a simple proof that the greedy algorithm indeed finds the best subspaces of every dimension.

Proposition

Let $A \in \mathbb{K}^{I \times J}$ and $v_1, v_2, \ldots, v_r$ be the singular vectors defined above. For $1 \leq k \leq r$, let V_k be the subspace spanned by $v_1, v_2, \ldots, v_k$. Then for each k, V_k is the best-fit k-dimensional subspace for A.

The natural oder of the singular values

We conclude this part with an important property left as an exercise: show that necessarily

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A), \tag{8}$$

i.e., the singular values come always with a natural nonincreasing order.

On matrix norms again

Note that the vector Av_i is really a list of lengths (with signs) of the projections of the rows of A onto v_i. Think of $\sigma_i(A) = \|Av_i\|_2$ as the component of the matrix A along v_i.

For this interpretation to make sense, it should be true that adding up the squares of the components of A along each of the v_i gives the square of the whole content of the matrix A.

This is indeed the case and is the matrix analogy of decomposing a vector into its components along orthogonal directions.

On matrix norms again

Consider one row, say $A^{(i)}$ of A. Since $v_1, v_2, \ldots, v_r$ span the space of all rows of A(exercise!), $\langle A^{(i)}, v\rangle = 0$ for all v orthogonal to $v_1, v_2, \ldots, v_r$.

Thus, for each row $A^{(i)}$, by Pythagoras Theorem, $\|A^{(i)}\|_2^2 = \sum_{k=1}^{r} |\langle A^{(i)}, v_k\rangle|^2$.

Summing over all rows $i \in I$ and recalling (6), we obtain

$$\sum_{i\in I} \|A^{(i)}\|_2^2 = \sum_{k=1}^{r}\sum_{i\in I} |\langle A^{(i)}, v_k\rangle|^2 = \sum_{k=1}^{r} \|Av_k\|_2^2 = \sum_{k=1}^{r} \sigma_k(A)^2.$$

But

$$\sum_{i\in I} \|A^{(i)}\|_2^2 = \sum_{i\in I}\sum_{j\in J} |A_{ij}|^2,$$

is the Frobenius norm of A, so that we obtained

$$\|A\|_F = \sqrt{\sum_{k=1}^{r} \sigma_k(A)^2}. \tag{9}$$

On matrix norms again

We might also observe that

$$\sigma_1(A) = \max_{\|v\|_2=1} \|Av\|_2 = \arg\max_{v\neq 0} \|Av\|_2/\|v\|_2 = \|A\|, \quad (10)$$

i.e., the first singular vector corresponds to the spectral norm of A. This allows us to define other and more general norms, called the Schatten-p-norms for matricesdefined for $1 \leq p < \infty$ by

$$\|A\|_p = \left(\sum_{k=1}^{r} \sigma_k(A)^p \right)^{1/p},$$

and for $p = \infty$

$$\|A\|_\infty = \max_{k=1,\ldots,r} \sigma_k(A).$$

Notice that these definitions gives precisely $\|A\|_F = \|A\|_2$ and $\|A\| = \|A\|_\infty$ as particular Schatten-norms. Of particular relevance in certain applications related to recommender systems is the so-called nuclear norm $\|A\|_* = \|A\|_1$ corresponding to the Schatten-1-norm.

On rank again

We just mentioned above, and left it as an exercise, that the orthogonal basis consitituted by the vectors $v_1, v_2, \ldots, v_r$ spans the space of all rows of A.

This means that the number r (of them) is actually the rank of the matrix! And this a proof of (g) in the Proposition above characterizing the rank!

The Singular Value Decomposition in its full glory

The vectors $Av_1, Av_2, \ldots, Av_r$ form yet another fundamental set of vectors associated with A: We normalize them to length one by setting

$$u_i = \frac{1}{\sigma_i(A)} Av_i$$

The vectors $u_1, u_2, \ldots, u_r$ are called the left singular vectors of A.

The v_i are called the right singular vectors.

The SVD theorem (soon!) will fully explain the reason for these terms.

Orthonormality of the left singular vectors

Clearly, the right singular vectors are orthogonal by definition.

We now show that the left singular vectors are also orthogonal.

Theorem

Let $A \in \mathbb{K}^{I \times J}$ be a rank r matrix. The left singular vectors of A, $u_1, u_2, \ldots, u_r$ are orthogonal.

SVD in its full glory

Theorem

Let $A \in \mathbb{K}^{I \times J}$ with right singular vectors $v_1, v_2, \ldots, v_r$, left singular vectors $u_1, u_2, \ldots, u_r$, and corresponding singular values $\sigma_1, \ldots, \sigma_r$. Then

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$$

Best rank-k approximation

Given the singular value decomposition $A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$ of a matrix $A \in \mathbb{K}^{I \times J}$, we define the k-rank truncation by

$$A_k = \sum_{\ell=1}^{k} \sigma_\ell u_\ell v_\ell^H.$$

It should be clear that A_k is a k-rank matrix.

Lemma

The rows of A_k are the orthogonal projections of the rows of A onto the subspace V_k spanned by the first k singular vectors of A.

Best rank-k Frobenius approximation

Theorem

For any matrix B of rank at most k

$$\|A - A_k\|_F \leq \|A - B\|_F$$

Best rank-k spectral approximation

Lemma

$\|A - A_k\|^2 = \sigma_{k+1}^2$.

Theorem

For any matrix B of rank at most k

$$\|A - A_k\| \leq \|A - B\|.$$

Nonconvexity and curse of dimensionality

We introduced the SVD by means of a greedy algorithm, which at each step is supposed to solve a nonconvex optimization problem of the type

$$v_k = \arg\max_{\|v\|_2=1, \langle v_1, v\rangle = 0, \ldots, \langle v_{k-1}, v\rangle = 0} \|Av\|_2. \tag{11}$$

Well, the issue with this is that it is neither a convex optimization problem nor, in principle, stated on a low dimensional space. In fact our goal is often to deal with "big data" (i.e., many data points, each of very high dimension), which eventually implies large dimensionalities of the matrix A.

How can we then approach the computation of the SVD without incurring in the so called "curse of dimensionality", which is an emphatic way of describing the computational infeasibility of a certain numerical task?

The Power Method

It is easiest to describe first in the case when A is real-valued, square, and symmetric and has the same right and left singular vectors, namely, $A = \sum_{k=1}^{r} \sigma_k v_k v_k^T$.

In this case we have

$$A^2 = \left(\sum_{k=1}^{r} \sigma_k v_k v_k^T \right) \left(\sum_{\ell=1}^{r} \sigma_\ell v_\ell v_\ell^T \right) = \sum_{k,\ell=1}^{r} \sigma_\ell \sigma_k v_k \underbrace{v_k^T v_\ell}_{:=\delta_{k\ell}} v_\ell^T = \sum_{k=1}^{r} \sigma_k^2 v_k v_k^T.$$

Similarly, if we take the m^{th} power of A, again all the cross terms are zero and we will get

$$A^m = \sum_{k=1}^{r} \sigma_k^m v_k v_k^T.$$

We had the spectral norm of A and a spectral gap

If we had $\sigma_1 \gg \sigma_2$, we would have

$$\lim_{m\to\infty} \frac{A^m}{\sigma_1^m} = v_1 v_1^T.$$

While it is not so simple to compute σ_1 (why?!), which corresponds to the spectral norm of A, one can easily compute the Frobenius norm $\|A^m\|_F$, so that we can consider $\frac{A^m}{\|A^m\|_F}$ which again converges for $m \to \infty$ to $v_1 v_1^T$ from which v_1 may be computed (exercise!).

What if A is not squared?

But then $B = AA^T$ is squared. If again, the SVD of A is $A = \sum_{k=1}^{r} \sigma_k u_k v_k^T$ then

$$B = \sum_{k=1}^{r} \sigma_k^2 u_k u_k^T.$$

This is the spectral decomposition of B. Using the same kind of calculation as above, one obtains

$$B^m = \sum_{k=1}^{r} \sigma_k^{2m} u_k u_k^T.$$

As m increases, for $k > 1$ the ratio $\sigma_k^{2m}/\sigma_1^{2m}$ goes to zero and B^m gets approximately equal to

$$\sigma_1^{2m} u_1 u_1^T,$$

provided again that $\sigma_1 \gg \sigma_2$. This suggests how to compute both σ_1 and u_1, simply by powering B.

Two relevant issues

- If we do not have a spectral gap $\sigma_1 \gg \sigma_2$ we get into troubles.
- Computing B^m costs m matrix-matrix multiplication, when done in the schoolbook way it costs $\mathcal{O}(md^3)$ operations or $\mathcal{O}(m \log d)$ when done by successive squaring.

Use a **random** vector application instead!

Instead, we compute

$$B^m x,$$

where x is a random unit length vector.

The idea is that the component of x in the direction of u_1, i.e., $x_1 := \langle x, u_1 \rangle$ is actually bounded away from zero with high probability and would get mutiplied by σ_1^2, while the other components of x along other u_i's would be multiplied by $\sigma_k^2 \ll \sigma_1^2$ only.

Each increase in m requires multiplying B by the vector $B^{m-1}x$, which we can further break up into

$$B^m x = A(A^T(B^{m-1}x)).$$

This requires two matrix vector products, involving the matrices A^T and A. Since $B^m x \approx \sigma_1^{2k} u_1(u_1^T x)$ is a scalar multiple of u_1, u_1 can be recovered from $B^m x$ by normalization.

Concentration of measure phenomenon

The concentration of measure phenomenon informally speaking is describing the fact that - in high-dimension - certain events happen with high-probability. In particular, we have

Lemma

Let $x \in \mathbb{R}^d$ be a unit d-dimensional vector of components $x = (x_1, \ldots, x_d)$ with respect to the canonincal basis and picked at radnom from the set $\{x : \|x\|_2 \leq 1\}$. The probability that $|x_1| \geq \alpha > 0$ is at least $1 - 2\alpha\sqrt{d-1}$.

Extensions

Remark

Notice that in the previous result essentially shows also that, independently of the dimension d, the $x_1 = \langle x, u_1 \rangle$ component of a random unit vector x with respect to any orthonormal basis $\{u_1, \ldots, u_d\}$[2] is bounded away from zero with overwhelming probability.

Remark

In view of the isometrical mapping $(a, b) \to a + ib$ from $\mathbb{R}^2$ to $\mathbb{C}$ the previous result extends to random unit vectors in $\mathbb{C}^d$ simply by modifying the statement as follows: The probability that, for a randomly chosen unit vecor $z \in \mathbb{C}^d$ $|z_1| \geq \alpha > 0$ holds is at least $1 - 2\alpha\sqrt{2d-1}$.

---

[2] Since the sphere is rotation invariant, the arguments above apply to any orthonormal basis by rotating it to coincide with the canonical basis!

Randomized Power Method

By injecting randomization in the process, we are able to use the blessing of the dimensionality (concentration of measure) against the curse of dimensionality.

Theorem

Let $A \in \mathbb{K}^{I \times J}$ and $x \in \mathbb{K}^{I}$ be a random unit length vector. Let V be the space spanned by the left singular vectors of A corresponding to singular values greater than $(1-\varepsilon)\sigma_1$. Let m be $\Omega\left(\frac{\ln(d/\varepsilon)}{\varepsilon}\right)$. Let w be unit vector after m iterations of the power method, namely,

$$w^* = \frac{(AA^H)^m x}{\|(AA^H)^m x\|_2}.$$

The probability that w^* has a component of at least $\mathcal{O}\left(\frac{\varepsilon}{\alpha d}\right)$ orthogonal to V is at most $1 - 2\alpha\sqrt{2d-1}$.