Orthogonal Projections

Vectors and Matrices

Ömür Uğur

Institute of Applied Mathematics
Middle East Technical University

RKHS (Learning) Seminars, March 2024

RKHS

Table of Contents

Continue with ...

Aims of this Talk

In fact, this is just a preliminary talk focusing on *orthogonal projections* using vectors and matrices.
However, we will also describe the *least-squares problem* in regression and try solving it in connection with orthogonal projection.

Yes or No

In order to achieve this we will

- *not* go in to details
- *not* prove (almost any) theorems, unless we will benefit from it later
- define some useful notations and notions
- define factorisation of positive (semi-) definite matrices

RKHS

Continue with . . .

Basic Notations and Definitions

- let $x, y \in \mathbb{R}^n$ be vectors
- define $\langle x, y\rangle = y^\top x = \sum_{i=1}^n y_i x_i$, and recall $\|x\|^2 = \langle x, x\rangle$
- define $\langle x, y\rangle = \|x\| \, \|y\| \cos\theta$
- $\langle x, y\rangle = 0$ then $\cos\theta = 0$ for $x, y \neq 0$; we say x and y are *orthogonal* and write $x \perp y$
- for a linear subspace $S \subset \mathbb{R}^n$ we call $x \in \mathbb{R}^n$ *orthogonal* to S if $x \perp z$ for all $z \in S$; and write it as $x \perp S$

Orthogonal Complement

Definition

The *orthogonal complement* of a linear subspace $S \subset \mathbb{R}^n$ is the set

$$S^{\perp} = \{x \in \mathbb{R}^n : x \perp S\}$$

Lemma (Orthogonal Complement)

$S^{\perp}$ is a linear subspace of $\mathbb{R}^n$

Proof.

Let $x, y \in S^{\perp}$ and $\alpha, \beta \in \mathbb{R}$, then for any $z \in S$ we have

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle = 0.$$

Hence $\alpha x + \beta y \in S^{\perp}$ □

Orthogonal Sets

Definition (Orthogonal–Orthonormal Sets)

A set of vectors, $\{x_1, \dots, x_k\} \subset \mathbb{R}^n$, is called an *orthogonal set* if, $x_i \perp x_j$ whenever $i \neq j$; it is called an *orthonormal set* if, in addition to orthogonallity, we have $\|x_i\| = 1$ for all i.

Pythagorean

Theorem (Pythagorean)

If $\{x_1, \dots, x_k\}$ *is an orthogonal set, then*

$$\|x_1 + \cdots + x_k\|^2 = \|x_1\|^2 + \cdots + \|x_k\|^2 .$$

Proof.

Particularly for $k = 2$, and $x_1 \perp x_2$, we have

$$\|x_1 + x_2\|^2 = \langle x_1 + x_2, x_1 + x_2 \rangle = \langle x_1, x_1 \rangle + 2 \langle x_1, x_2 \rangle + \langle x_2, x_2 \rangle$$

Hence, $\|x_1 + x_2\|^2 = \|x_1\|^2 + \|x_2\|^2$. □

Linear Independence

If $X \subset \mathbb{R}^n$ is an orthogonal and $0 \notin X$, then X is *linearly independent*.

Continue with . . .

Orthogonal Projection

Theorem (Orthogonal Projection)

Let $x \in \mathbb{R}^n$ *and* $S \subset \mathbb{R}^n$ *be a linear subspace, there exists a unique solution to the minimisation problem*

$$\hat{x} = \arg\min_{z \in S} \|x - z\| .$$

The minimiser $\hat{x}$ *is the unique vector in* $\mathbb{R}^n$ *that satisfies*

$$\hat{x} \in S \quad \text{and} \quad x - \hat{x} \perp S.$$

The vector $\hat{x}$ *is called the orthogonal projection of* x *onto* S*.*

RKHS

Proof of Orthogonal Projection

The proof here contains only the *sufficiency*.

Proof of Orthogonal Projection Theorem.

Let $x \in \mathbb{R}^n$ and $S \subset \mathbb{R}^n$ be a linear subspace. Let $\hat{x} \in S$ such that $x - \hat{x} \perp S$. Let $z \in S$ be any (other) vector in S, then by the Pythagorean theorem, we have

$$\|x - z\|^2 = \|(x - \hat{x}) + (\hat{x} - z)\|^2 = \|x - \hat{x}\|^2 + \|\hat{x} - z\|^2$$

Hence, $\|x - z\| \geq \|x - \hat{x}\|$ for all $z \in S$. □

What $\hat{x} \in S$ best approximates a given $x \in \mathbb{R}^n$?

Projection as a Mapping

By the Orthogonal Projection Theorem, there is a well-defined mapping (or operator from $\mathbb{R}^n$ to $\mathbb{R}^n$)

$$x \in \mathbb{R}^n \mapsto \text{its orthogonal projection } \hat{x} \in S \subset \mathbb{R}^n.$$

We denote this operator by P and let $\hat{x} = Px$ and call it *orthogonal projection operator*. Some uses the notation, $\hat{E}_S x = Px$, where $\hat{E}_S$ is called *wide-sense expectation operator*.

Thus from the Orthogonal Projection Theorem, we have

- $Px \in S$ and $x - Px \perp S$
- $\|x\|^2 = \|Px\|^2 + \|x - Px\|^2$
- $\|Px\| \leq \|x\|$

The second item above simply follows from writing

$$x = \underbrace{Px}_{\in S} + \underbrace{x - Px}_{\in S^\perp},$$

and applying Pythagorean theorem.

Orthogonal Complement

Let X be a linear subspace with linear subspace S and its complement $S^\perp$; we write

$$X = S \oplus S^\perp$$

to indicate that for every $x \in X$, there is a unique $x_1 \in S$ and a unique $x_2 \in S^\perp$ such that

$$x = x_1 + x_2.$$

Moreover,

$$x_1 = \hat{E}_S x \quad \text{and} \quad x_2 = x - \hat{E}_{S^\perp} x.$$

Theorem (Another Version of Orthogonal Projection)

If S is a linear subspace of $\mathbb{R}^n$, $\hat{E}_S x = Px$ and $\hat{E}_{S^\perp} x = Mx$, then for any $x \in \mathbb{R}^n$

$$Px \perp Mx \quad \textit{and} \quad x = Px + Mx.$$

Continue with ...

Orthonormal Basis

Recall that an orthogonal set of vectors $\mathcal{O} \subset \mathbb{R}^n$ is called an *orthonormal* set if $\|u\| = 1$ for all $u \in \mathcal{O}$.

Definition (Orthogonal Basis)

Let S be a linear subspace of $\mathbb{R}^n$ and $\mathcal{O} \subset S$; if $\mathcal{O}$ is orthonormal and $\operatorname{span} \mathcal{O} = S$, then $\mathcal{O}$ is called an *orthonormal basis* for S.

Recall that $\mathcal{O}$ is necessarily a *basis*, since the vectors in an orthogonal set are linearly independent.

RKHS

Orthonormal Basis

Lemma

If $\{u_1, \ldots, u_k\}$ *is an orthonormal basis for a linear subspace* S*, then for any* $x \in S$*,*

$$x = \sum_{i=1}^{k} \langle x, u_i \rangle \, u_i.$$

Proof.

Since $x \in \operatorname{span}\{u_1, \ldots, u_k\}$, there are α_j such that $x = \sum_{j=1}^{k} \alpha_j u_j$. Therefore,

$$\langle x, u_i \rangle = \sum_{j=1}^{k} \alpha_j \langle u_j, u_i \rangle = \alpha_i,$$

due the orthogonality that $\langle u_j, u_i \rangle = \delta_{ij}$. □

Projection onto Orthonormal Basis

Theorem (Projection)

If $\mathcal{O} = \{u_1, \ldots, u_k\}$ *is an orthonormal basis for a linear subspace* S, *then for any* $x \in \mathbb{R}^n$,

$$Px = \sum_{i=1}^{k} \langle x, u_i \rangle \, u_i.$$

Proof.

Consider Px above; clearly, $Px \in S$ and for any $u_j \in \mathcal{O}$,

$$\left\langle x - \sum_{i=1}^{k} \langle x, u_i \rangle \, u_i, u_j \right\rangle = \langle x, u_j \rangle - \sum_{i=1}^{k} \langle x, u_i \rangle \, \langle u_i, u_j \rangle$$
$$= \langle x, u_j \rangle - \langle x, u_j \rangle = 0.$$

Hence, $x - Px \perp S$. □

Continue with . . .

Projection using Matrices

We have already mentioned that the projection can be considered a linear mapping from $x \in \mathbb{R}^n$ to $Px \in \mathbb{R}^n$.

Theorem (Projection Matrix)

Let the columns of an $n \times k$ matrix X form a basis for S. Then,

$$P = X(X^\top X)^{-1}X^\top$$

is a projection (matrix) onto S.

Remark. The matrix $M = I - P$ satisfies $Mx = \hat{E}_{S^\perp}x$ and it is sometimes called an *annihilator* matrix.

Proof of Projection Matrix.

Let $x \in \mathbb{R}^n$ and $P = X(X^\top X)^{-1}X^\top$.

- $Px = X \underbrace{(X^\top X)^{-1}X^\top x}_{a} = Xa$, hence $Px \in S$.
- Notice that for any $y \in \mathbb{R}^k$, we have $z = Xy \in S$ and using $\langle x - Px, z\rangle = z^\top(x - Px)$, we calculate

$$\underbrace{(Xy)^\top}_{z^\top}\underbrace{\left[x - X(X^\top X)^{-1}X^\top x\right]}_{x-Px} = y^\top \left[X^\top x - \underbrace{X^\top X(X^\top X)^{-1}}_{I_k} X^\top x\right] = 0.$$

So, $x - Px \perp S$.

Hence, the proof is completed. □

Corollary for Orthonormal Case

Theorem (Corollary for Orthonormal Case)

Suppose U is an $n \times k$ matrix with orthonormal columns; let $u_i = \text{col}_i U$ and let $S = \operatorname{span} U = \operatorname{span}\,\{u_1, \ldots, u_k\}$. Then,

$$P = UU^\top \quad \textit{and} \quad Px = UU^\top x = \sum_{i=1}^{k} \langle x, u_i \rangle \, u_i.$$

Proof. Since the columns of U are orthonormal, it satisfies $U^\top U = I_k$; thence,

$$P = U(U^\top U)^{-1} U^\top = UU^\top.$$

The final part of the theorem, that is, $Px = \sum_{i=1}^{k} \langle x, u_i \rangle \, u_i$, directly follows from the Projection Theorem (above); to recall and check:

$$\alpha_i = \langle Px, u_i \rangle = u_i^\top UU^\top x = u_i^\top x = \langle x, u_i \rangle \, .$$

Continue with ...

RKHS

Over-determined Systems

- Let $y \in \mathbb{R}^n$ and X be an $n \times k$ matrix with linearly independent columns; we seek a vector (of unknowns) $b \in \mathbb{R}^k$ satisfying $Xb = y$.
- If $n > k$ (i.e., more equations than unknowns) then b (or the system) is to be *over-determined*.
- And, in general, we seek for an approximate solution: $b \in \mathbb{R}^k$ such that Xb is close to y. Such a solution is well-defined and unique.

Theorem

The unique minimiser of $\|y - Xb\|$ over $b \in \mathbb{R}^k$ is $\hat{\beta} = (X^\top X)^{-1} X^\top y$.

Proof. Note that $X\hat{\beta} = X(X^\top X)^{-1} X^\top y = Py$, that is, Py is an orthogonal projection onto $\operatorname{span} X$; thence,

$$\|y - Py\| \leq \|y - z\| \quad \text{for all } z \in \operatorname{span} X.$$

Particularly, since $Xb \in \operatorname{span} X$ and $Py = X\hat{\beta}$,

$$\left\| y - X\hat{\beta} \right\| \leq \|y - Xb\| \quad \text{for all } b \in \mathbb{R}^k.$$

Least-Squares Regression

Given the pairs $(x, y) \in \mathbb{R}^k \times \mathbb{R}$, and let $f : \mathbb{R}^k \to \mathbb{R}$ in order to minimise the risk (loss)

$$\mathcal{R}(f) = \mathbb{E}\left[(y - f(x))^2\right].$$

- Unless the underlying probability or the expectation is given, we cannot solve the problem!
- However, if a *sample* of size n is provided, we can *estimate* the risk: *empirical risk*:

$$\operatorname*{minimise}_{f \in \mathcal{F}} \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

However, still this includes *calculus of variations* — generally, $\mathcal{F}$ is called a *hypothesis space* and is suggested to be *simple* (to avoid over fitting).

RKHS

Linear Least-Squares (sample problem)

Let $\mathcal{F}$ be the class of (all) linear functions defined as

$$\mathcal{F} = \left\{ f : f(x) = b^\top x, \quad x \in \mathbb{R}^k \right\}.$$

Thus the problem is

Definition (Linear Least-Squares)

$$\underset{b \in \mathbb{R}^k}{\text{minimise}}\ \hat{R}(f) = \sum_{i=1}^{n} \left(y_i - b^\top x_i \right)^2.$$

Define: $y = [y_1, \ldots, y_n]^\top$, $x_i = [x_{i1}, \ldots, x_{ik}]^\top$ and

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}.$$

We assume that $n > k$ and X has a full column rank.

Linear Least-Squares as Projection Problem

With the notations introduced, an easy algebra shows that

$$\|y - Xb\|^2 = \sum_{i=1}^{n} \left(y_i - b^\top x_i \right)^2$$

and since a monotone transformation does not effect the minimisers, the least-squares problem turns into

$$\arg\min_{b \in \mathbb{R}^k} \sum_{i=1}^{n} \left(y_i - b^\top x_i \right)^2 = \arg\min_{b \in \mathbb{R}^k} \|y - Xb\| \, .$$

Thence, the solution (by the over-determined system results):

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

RKHS

Linear Least-Squares as Projection Problem

- let P and M be the projection and annihilator associated with X:

$$P = X(X^\top X)^{-1}X^\top \quad \text{and} \quad M = I - P.$$

- The vector of *fitted* values is $\hat{y} = X\hat{\beta} = Py$
- The vector of *residuals* is $\hat{r} = y - \hat{y} = y - Py = My$

Here are some standard definitions (and a theorem):

- TSS $= \|y\|^2$ (total sum of squares)
- SSR $=$ RSS $= \|r\|^2$ (sum of squared residuals)
- ESS $= \|\hat{y}\|^2$ (explained sum of squares)

Theorem (TSS = ESS + SSR)

$$TSS = ESS + SSR$$

Proof. $y = \hat{y} + \hat{r}$ and $\hat{r} \perp \hat{y}$, then by Pythagorean theorem

$$\|y\|^2 = \|\hat{y} + \hat{r}\|^2 = \|\hat{y}\|^2 + \|\hat{r}\|^2 .$$

RKHS

Continue with . . .

Gram-Schmidt Orthogonalisation

Theorem (Orthonormal Basis)

For linearly independent set $\{x_1, \ldots, x_k\} \subset \mathbb{R}^n$*, there is an orthonormal set* $\{u_1, \ldots, u_k\} \subset \mathbb{R}^n$ *with*

$$\operatorname{span}\{x_1, \ldots, x_i\} = \operatorname{span}\{u_1, \ldots, u_i\}, \qquad i = 1, \ldots, k.$$

Gram-Schmidt Orthogonalisation

Gram-Schmidt Orthogonalisation Procedure

1. for $i = 1, \ldots, k$ form $S_i = \operatorname{span}\{x_1, \ldots, x_i\}$ and $S_i^\perp$
2. set $v_1 = x_1$
3. for $i \geq 2$, set $v_i = \hat{E}_{S_{i-1}^\perp} x_i$ and $u_i = \frac{v_i}{\|v_i\|}$

Equivalently, as commonly appears: set $v_1 = x_1$, then for $i = 2, \ldots, k$,

$$v_i = x_i - \sum_{j=1}^{i-1} \operatorname{proj}_{v_j} x_i,$$

where

$$\operatorname{proj}_v x = \frac{\langle x, v \rangle}{\langle v, v \rangle} v,$$

and consequently,

$$u_i = \frac{v_i}{\|v_i\|}.$$

QR Decomposition

Theorem (QR Decomposition)

If X is an $n \times k$ matrix with linearly independent columns, then there exists a factorisation of the form $X = QR$ where

- *R is $k \times k$, upper triangular and nonsingular;*
- *Q is $n \times k$ with orthonormal columns.*

Proof (sketch only). Let $x_j = \text{col}_j X$, and let $\{u_1, \ldots, u_k\}$ be the orthonormal set with the same span of $\{x_1, \ldots, x_k\}$, by Gram-Schmidt process (for instance).
Let $Q = [u_1, \ldots, u_k]$ be the matrix with columns u_i. Then, since $x_j \in \operatorname{span}\{u_1, \ldots, u_j\}$, we have $j = 1, \ldots, k$:

$$x_j = \sum_{i=1}^{j} \langle x_j, u_i \rangle \, u_i, \quad \text{equivalently} \quad X = QR.$$

Linear Regression with QR Decomposition

We have seen that the over-determined system, $Xb = y$, has the least-squares approximation as

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Using this and the decomposition $X = QR$, we get

$$\begin{aligned}\hat{\beta} &= (R^\top Q^\top QR)^{-1} R^\top Q^\top y = (R^\top R)^{-1} R^\top Q^\top y \\ &= R^{-1} R^{-\top} R^\top Q^\top y \\ &= R^{-1} Q^\top y.\end{aligned}$$

Hence, the solution blows down to *back-substitution* in

$$R\hat{\beta} = Q^\top y.$$

RKHS

Continue with . . .

Positive Definite Matrices

Definition (Positive Definite Matrices)

A matrix A is *positive definite* if

$$\langle Ax, x \rangle = x^\top Ax > 0$$

for all nonzero x.

A positive definite matrix

- has *real* and *positive* eigenvalues,
- its leading principal submatrices all have positive determinants
- has positive diagonal elements.

Cholesky Decomposition

Theorem (Cholesky Decomposition)

A Cholesky decomposition,

$$A = UU^\top,$$

of A, where U is an upper triangular matrix, exists if, and only if, A is symmetric and positive definite.

Definition (Square Root Decomposition)

A *square root* of a matrix A is defined as a matrix S such that

$$S^2 = SS = A.$$

Generaly, we use the notation $A^{1/2}$ instead of S.

Eigendecomposition – Spectral Decomposition

Theorem (Spectral Decomposition)

Let A be an $n \times n$ matrix with n linearly independent eigenvectors, say u_i corresponding to λ_i. Then A has the spectral decomposition,

$$A = U \Lambda U^{-1},$$

where U is the square $n \times n$ matrix whose ith column is the eigenvector u_i of A, and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues λ_i: $U = [u_1, \dots, u_n]$ and $\Lambda_{ii} = \lambda_i$.
In most cases, the normalised eigenvectors u_i are chosen, but this is not necessary.

Square Root of an SPD Matrix

Hence, as a corollary, for a symmetric positive definite matrix A, we have the eigendecompostion as

$$A = U\Lambda U^\top,$$

where, in this case, U is an orthogonal matrix whose columns are the orthonormalised eigenvectors of A. In other words, we can choose an *orthonormal* set of eigenvectors u_i. Such a statement needs a proof though!

Hence, for such a symmetric positive definite matrix A, we have

$$A = U\Lambda U^\top = \left(U\Lambda^{1/2}U^\top\right)\left(U\Lambda^{1/2}U^\top\right) = SS,$$

so that

$$A^{1/2} = U\Lambda^{1/2}U^\top.$$

RKHS

Continue with ...

Singular Value Decomposition

Definition

An $m \times n$ matrix A has (always) a *singular value decomposition* of the form

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

where U $(m \times m)$ and V $(n \times n)$ are orthogonal (respectively, the left and right singular vector) matrices and Σ is a diagonal one, containing the (non-negative) singular values.

Singular Value Decomposition

Particularly,

- if $m \geq n$, then

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^\top, \qquad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}$$

  with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

- if $m \leq n$, then

$$A = U \begin{bmatrix} \Sigma & 0 \end{bmatrix} V^\top, \qquad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}$$

  with $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$.

Singular Value Decomposition - Consequences

Apart from many properties, it is important to recall the following three; Let $A = U\Sigma V^\top$ with $U = [u_1, \ldots, u_p]$, $V = [v_1, \ldots v_p]$ and $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_p\}$, where $p = \min\{m, n\}$ and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.

- For such an A, the singular vectors satisfy

$$Av_i = \sigma_i u_i, \qquad A^\top u_i = \sigma_i v_i,$$

  equivalently,

$$A^\top A v_i = \sigma_i^2 v_i, \qquad AA^\top u_i = \sigma_i^2 u_i,$$

  for $1 \leq i \leq p$.

- If $r = \operatorname{rank} A$, then

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^\top.$$

- Finally, A is symmetric positive definite (square) matrix if, and only if, its singular value decomposition is $A = V\Sigma V^\top$, where Σ is nonsingular.

RKHS

Continue with ...

Complex Matrices

When the entries of a matrix are in $\mathbb{C}$: $z = \alpha + i\beta$, $\overline{z} = \alpha - i\beta$ and $|z| = \sqrt{\alpha^2 + \beta^2}$, where $i = \sqrt{-1}$, here are the *Real vs Complex Correspondence*.

- transpose vs conjugate (Hermitian) transpose

$$A^\top \quad \text{versus} \quad A^* = A^{\mathrm{H}} = \overline{A^\top}$$

- symmetric vs Hermitian

$$A^\top = A \quad \text{versus} \quad A^{\mathrm{H}} = A$$

  In this case, we call A, sometimes, *self-adjoint*.

- orthogonal vs unitary

$$A^\top A = AA^\top = I \quad \text{versus} \quad A^{\mathrm{H}} A = AA^{\mathrm{H}} = I$$

  That is, when the inverse $A^{-1} = A^{\mathrm{H}}$.

Bibliography

Jonathan H. Manton and Pierre-Olivier Amblard.
A primer on reproducing kernel Hilbert spaces, 2015.

Il Shan Ng.
Reproducing kernel Hilbert spaces & machine learning, 2024.
accessed: March 2024.

Vern I. Paulsen and Mrinal Raghupathi.
An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.
Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.

Jesse Perla, Thomas J. Sargent, and John Stachurski.
Orthogonal projections and their applications, 2024.
accessed: March 2024.

RKHS