MATH 462 LECTURE NOTES
INNER PRODUCTS AND PROJECTIONS

ADAM M. OBERMAN

1. Inner Products

1.1. **Review of analytic geometry.** Review [DFO20, Chapter 3], sections 3.1-3.6

- Definition of norms (normed vector space), 1-norm, 2-norm
- Definition of inner products (inner product space)
- Definition of PSD (symmetric, positive definite) matrix
- Definition of a metric
- Cauchy Schwartz inequality
- Angle between two vectors: $\cos\theta = x^\top y/\|x\|\|y\|$.

1.2. **Inner Products.** Given two vectors, $x, z \in \mathbb{R}^d$, the inner product is

$$x \cdot z = x^\top z = \sum_{i=1}^{d} x_i z_i$$

Definition 1.1. Let V be a vector space and let $B : V \times V \to \mathbb{R}$ be a bilinear mapping that takes two vectors and maps them onto a real number. Then

- B is called symmetric if $B(x, y) = B(y, x)$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$,
- B is called non-negative definite if

$$B(x, x) \geq 0, \qquad \forall x \in V,$$

- B is called positive definite, if it is non-negative definite, and if, in addition,

$$B(x, x) = 0 \iff x = 0$$

A positive definite, symmetric bilinear mapping $B : V \times V \to \mathbb{R}$ is called an inner product on V. We typically write $\langle x, y\rangle$ instead of $B(x, y)$. The pair $(V, \langle\cdot,\cdot\rangle)$ is called an inner product space.

Example 1.2. Given two vectors, $x, z \in \mathbb{R}^d$, we can define a different inner product than the usual one. For example, if we have positive vector $a = (a_1, \ldots, a_d)$ and rescale the variables,

$$\tilde{x} = (a_1 x_1, \ldots, a_d x_d)$$

then we can define an inner product

$$\langle x, z\rangle_a = \tilde{x} \cdot \tilde{z} = (Mx) \cdot (Mx), \qquad M = diag(a)$$

More generally, given any full rank matrix, M, we can define

$$\langle x, z\rangle_M = (Mx) \cdot (Mz),$$

Exercise 1.1. Prove that if M is full rank, then $\langle x, z\rangle_M = (Mx) \cdot (Mx)$ is an inner product. Why does M have to be full rank? Hint: show that if $Mx = 0$ for some nonzero x, then the positive definite requirement fails.

Date: October 24, 2023.

Example 1.3. Let $\lambda \in \mathbb{R}^d$. Define $B(x,z) = x^\top diag(\lambda) z$ Then B is bilinear. Note if any component of λ is negative, say λ_j, then $B(e_j, e_j) = \lambda_j < 0$.

Exercise 1.2. Let $\lambda \in \mathbb{R}^d$, and let M be a full rank $d \times d$ matrix. Define $B(x,z) = (Mx)^\top diag(\lambda) Mz$. Show that B is a symmetric bilinear form. Show that B is non-negative definite if and only if each component λ is non-negative. Likewise, show that B is positive definite if and only if each component of λ is positive.

1.3. **Symmetric PSD matrices.**

Definition 1.4. A symmetric matrix, P, is called non-negative definite if $x^T P x \geq 0$ for all x. A non-negative definite matrix is called positive definite if $x^T P x = 0$ only when $x = 0$.

Example 1.5. Let M be an $d \times d$ matrix.

Define $P = M^\top M$.

Then P is non-negative definite. If, in addition, M is full rank, then P is positive definite.

Example 1.6. For example, set $M = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}$ Then $P = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 9 \end{bmatrix}$. The matrix P is non-negative definite. Note $x^\top P x = (Mx)^\top (Mx)$ which is non-negative, since it is $\|v\|^2$ for $v = Mx$. An additional argument shows P is positive definite.

Example 1.7 (from textbook). Consider the matrices

$$\boldsymbol{A}_1 = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 9 & 6 \\ 6 & 3 \end{bmatrix}.$$

$\boldsymbol{A}_1$ is positive definite because it is symmetric and

$$\begin{aligned} \boldsymbol{x}^\top \boldsymbol{A}_1 \boldsymbol{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 9x_1^2 + 12 x_1 x_2 + 5 x_2^2 = (3x_1 + 2x_2)^2 + x_2^2 > 0 \end{aligned}$$

for all $\boldsymbol{x} \in V \backslash \{\mathbf{0}\}$. In contrast, $\boldsymbol{A}_2$ is symmetric but not positive definite because $\boldsymbol{x}^\top \boldsymbol{A}_2 \boldsymbol{x} = 9x_1^2 + 12 x_1 x_2 + 3 x_2^2 = (3x_1 + 2x_2)^2 - x_2^2$ can be less than 0, e.g., for $\boldsymbol{x} = [2, -3]^\top$.

Remark 1.8. In the two dimensional example above, the proof was done by completing the square. This is possible in two dimensions. However, the more general higher dimensional case is not so easy. Instead, we have to rely on advanced linear algebra techniques.

1.4. **Orthogonal Matrices and Matrix Factorizations.**

In order to determined if a matrix is positive definite, we need to use advanced linear algebra techniques, in particular, matrix factorizations. `https://en.wikipedia.org/wiki/Matrix_decomposition`

Two vectors v, z are called orthogonal if

$$v \cdot z = 0$$

The vector v is a unit vector if

$$\|v\|^2 = v \cdot v = 1$$

A square matrix O is called orthonomal if each row of the matrix consists of unit vectors, which are orthogonal. So the rows form an orthogonal basis of the $V = \mathbb{R}^d$.

For every $d \times d$ real symmetric matrix, the eigenvalues are real and the eigenvectors can be chosen real and orthonormal. Thus a real symmetric matrix $\mathbf{A}$ can be decomposed as

$$A = O^\top \Lambda O$$

where $\mathbf{Q}$ is an orthogonal matrix whose columns are the real, orthonormal eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$[7]

For example, if a matrix is PSD, then it can be factored as $P = O^\top \Lambda O$, where O orthogonal and Λ is diagonal with positive numbers on the diagonal.

More generally, any real symmetric matrix, A, can be decomposed as

$$A = O^\top \Lambda O$$

where Λ is diagonal, but can have zero or negative values on the diagonal.

1.5. **Matrix Factorizations and Outer product.** Note when we write

$$A = O^\top \Lambda O$$

Let the row (column?) vectors of O be denotes $v_1, \ldots, v_d$.

We can also write A as

$$A = \sum_{i=1}^{d} \lambda_i v_i v_i^\top$$

so that

$$Ax = \sum_{i=1}^{d} \lambda_i (v_i \cdot x) v_i$$

Example 1.9. if $x = [1, 2, 3]$ then

$$x^\top x = 1^2 + 2^2 + 3^2 = 14$$

but

$$xx^\top = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

Exercise 1.3. TODO: two dimensional example of this factorization.

2. Orthogonal Projections

Review [DFO20, Chapter 3], Section 3.8

- orthogonal vectors
- orthogonal projections
- projections onto line
- projections onto subspace
- projection matrices
- PSD Matrix factorization, $P = O^\top \Lambda O$, where O orthogonal and Λ is diagonal.

The projection onto a set S is defined to be

$$P(x) = \arg\min_{s \in S} \|x - s\|^2$$

(when it is unique).

The projection is unique for any convex set, including linear and affine subspaces.

2.1. **Projection onto a line.** A line in $\mathbb{R}^d$, in the direction $b \in \mathbb{R}^d$, is determined by the parametric equation

$$L = \{bt \mid t \in \mathbb{R}\}$$

The projection onto the line is defined by

$$P(x) = \arg\min_t \|x - bt\|^2$$

Minimizing this leads to

$$b^\top (bt - x) = 0$$

or

$$t = \frac{b \cdot x}{\|b\|^2}, \qquad P(x) = bt = \frac{b \cdot x}{\|b\|^2} b$$

We can write the projection as the linear operator

$$P = \frac{1}{\|b\|^2} bb^\top$$

We can write the matrix representation as of the projection as

$$M = \mathrm{Proj}_b = \frac{1}{\|b\|^2} bb^\top$$

Example 2.1. Example 3.10 (Projection onto a Line) Find the projection matrix $\boldsymbol{P}_\pi$ onto the line through the origin spanned by $\boldsymbol{b} = \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}^\top$. Here $\boldsymbol{b}$ is a direction and a basis of the one-dimensional subspace (line through origin). We obtain

$$\boldsymbol{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\boldsymbol{b}^\top \boldsymbol{b}} = \frac{1}{9} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 4 \end{bmatrix}.$$

Let us now choose a particular $\boldsymbol{x}$ and see whether it lies in the subspace spanned by $\boldsymbol{b}$. For $\boldsymbol{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$, the projection is

$$\pi_U(\boldsymbol{x}) = \boldsymbol{P}_\pi \boldsymbol{x} = \frac{1}{9} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 5 \\ 10 \\ 10 \end{bmatrix} \in \mathrm{span} \left[\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \right].$$

Note that the application of $\boldsymbol{P}_\pi$ to $\pi_U(\boldsymbol{x})$ does not change anything, i.e., $\boldsymbol{P}_\pi \pi_U(\boldsymbol{x}) = \pi_U(\boldsymbol{x})$. This is expected because according to Definition 3.10, we know that a projection matrix $\boldsymbol{P}_\pi$ satisfies $\boldsymbol{P}_\pi^2 \boldsymbol{x} = \boldsymbol{P}_\pi \boldsymbol{x}$ for all $\boldsymbol{x}$.

Remark 2.2. We can show that $\pi_U(\boldsymbol{x})$ is an eigenvector of $\boldsymbol{P}_\pi$, and the corresponding eigenvalue is 1 .

Definition 2.3. Given $x \in \mathbb{R}^n$ and a linear subspace U, we define the projection

$$\mathrm{Proj}_U(x) = \arg\min_{y \in U} \|x - y\|^2 \tag{V}$$

This is the *variational* definition of the projection, as the closest point.

When U has a basis $b_1, \ldots, b_p$, we can write the projection in the parametric form. Since any vector $y \in U$ can be written as

$$y = \sum_{i=1}^p \lambda_i b_i = B\lambda, \qquad B = [b_1, \ldots, b_p],\ \lambda \in \mathbb{R}^p$$

Then (V) is equivalent to

$$\text{(P)} \qquad \mathrm{Proj}_U(x) = \arg\min_{\lambda \in \mathbb{R}^p} \|B\lambda - x\|^2$$

which we refer to as the parametric representation.

Reviewing vector calculus rules as above (which use math notation). Now returning to ML notation, define $f(\lambda) = \|B\lambda - x\|^2$, then

$$\nabla_\lambda f(\lambda) = 2B^\top (B\lambda - x)$$

so the minimizer, λ, of (P) solves

$$\text{(1)} \qquad B^\top B\lambda = B^\top x$$

Here (1) is called the *normal equation*. Then $y = B\lambda$ gives

$$\text{(L)} \qquad \mathrm{Proj}_U(x) = B(B^\top B)^{-1}B^\top x$$

We refer to (L) as the matrix representation of the projection. In particular,

$$\mathrm{Proj}_U = B(B^\top B)^{-1}B^\top$$

2.2. **Orthogonal Basis.** If we use an orthonormal basis $v_1, \ldots, v_p$, and write

$$O = [v_1, \ldots, v_p]^\top, \quad p \times n \text{ matrix}$$

Then $O^T O = I$ is the p dimensional identity matrix, and (L) becomes

$$\mathrm{Proj}_U(x) = OO^\top x$$

Remark 2.4. See examples in class or from [DFO20] of orthogonal projection matrices.

Here we see that

$$M = \mathrm{Proj}_U = \sum_{i=1}^{p} \mathrm{Proj}_{v_i} = \sum_{i=1}^{p} v_i {v_i}^\top$$

which represents the projection matrix as a sum of one dimensional projections.

Example 2.5. Let U be the span of two vectors, $b_1 = [1,1,1]^\top$, $b_2 = [0,1,2]^\top$ in $\mathbb{R}^3$. Then ..., the projection matrix is given in notes.

Form, using Gram-Schmidt, the orthonormal basis $v_1 = \frac{1}{\sqrt{3}}[1,1,1]^\top$, $v_2 = \frac{1}{\sqrt{2}}[-1,0,1]^\top$. Then the projection matrix can be written as

$$M = \mathrm{Proj}_U = v_1 v_1^\top + v_2 v_2^\top = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

3. Principal components analysis

Refer to [DFO20] Chapter 10. Refer to [SSBD14], Chapter 23 for proofs.
Given $S^m = \{x_1, \ldots x_m\}$ with $x_i \in \mathbb{R}^n$.

Definition 3.1. The covariance matrix of S^n is given by

$$C = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^\top$$

Recall that $M = xx^\top$ is the rank 1 $n \times n$ matrix

$$M_{ij} = x_i x_j.$$

The vector representation. Given S^m as above, form the $m \times d$ matrix

$$X = [x_1, \ldots, x_m]^\top \in \mathbb{R}^{m \times d}$$

and write

$$X^\top = [x_1^\top, \ldots, x_m^\top] \in \mathbb{R}^{d \times m}$$

Then the covariance matrix is given by the $d \times d$ matrix

$$C = X^\top X \in \mathbb{R}^{d \times d}$$

Where

$$C = \sum_{i=1}^m x_i x_i^\top$$

(which follows from the matrix representations above).

Definition 3.2. Given S^m with covariance matrix C. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ be the non-negative eigenvalues of C and let $v_1, \ldots, v_n$ be the corresponding eigenvectors. Then the first p principal components are given by $v_1, \ldots, v_p$. Given a data point x, the PCA representation of x is given by the projection onto the span of $v_1, \ldots, v_p$

$$\mathrm{Proj}_V(x) = \sum_{i=1}^p \mathrm{Proj}_{v_i}(x) = \sum_{i=1}^p (v_i^\top x) v_i$$

We have the following variational interpretation of PCA.

Definition 3.3. (Compression and recovery matrix) Let W be a compression matrix mapping the data, vectors in $\mathbb{R}^n$ to $\mathbb{R}^p$, for $p < n$. Let U be a recovery matrix, mapping $\mathbb{R}^p$ to $\mathbb{R}^n$. For a given dataset S^m, with mean zero, define

$$L(W, U, S^m) = \frac{1}{m} \sum_{i=1}^m \|x_i - UWx_i\|^2 \tag{2}$$

Theorem 3.4. *Given S^m, then the Compression-Recovery loss* (2) *is minimized by $W = V$ and $U = V^\top$, where V is the matrix of the first p eigenvectors of the covariance matrix of the data.*

Proof. This theorem is proved in [SSBD14], Chapter 23. See also Calder notes. □

References

[DFO20] Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.