MATH 462 LECTURE NOTES

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: $\cos\theta = x^\top y/\|x\|\|y\|$.

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.

Example 2.1. Do all the examples in Section 3.8

Example 2.2. if $x = [1, 2, 3]$ then

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

but

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

2.1. **Projection onto vectors.** Given a vector b, the projection of x onto b is given by

$$\mathrm{Proj}_b(x) = \arg\min_t \|x - tb\|^2$$

Define $f(t) = \|x - tb\|^2$ so that $f'(t) = b^\top(x - tb)$, giving

$$t = \frac{b^\top x}{\|b\|^2}, \qquad tb = \frac{b^\top x}{\|b\|^2} b$$

Thus

$$\mathrm{Proj}_b(x) = \frac{b^\top x}{\|b\|^2} b$$

Date: November 8, 2022.

We can write the matrix representation as of the projection as

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

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

$$\mathrm{Proj}_U(x) = \operatorname*{arg\,min}_{y \in U} \|x - y\|^2 \tag{V}$$

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

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

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

Then (V) is equivalent to

$$\mathrm{Proj}_U(x) = \operatorname*{arg\,min}_{\lambda \in \mathbb{R}^p} \|B\lambda - x\|^2 \tag{P}$$

which we refer to as the parametric representation.

Remark 2.4 (Vector calculus review). Recall from vector calculus, `https://en.wikipedia.org/wiki/Gradient`.

(1) x is a d-dimensional column vector,
(2) $f : \mathbb{R}^d \to R$, Then $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$, $\nabla f(x)$ is also a column vector. The reason for this is we want to generalize the derivative: $f(x + h) \approx f(x) + hf'(x)$ becomes:

$$f(x + hv) \approx f(x) + h\nabla f(x) \cdot v$$

. We can't write the equation above if ∇f is a row vector.
(3) (The total derivative $df = \nabla f^\top$ is a row vector, see, `https://en.wikipedia.org/wiki/Gradient` total derivative.)
(4) If $g : \mathbb{R}^d \to \mathbb{R}^n$ (the function is a column vector), then the jacobian, $Jg : \mathbb{R}^d \to \mathbb{R}^n$, is the matrix of partial derivatives,

$$(Jg)_{ij} = \frac{\partial g_i}{\partial x_j}$$

Each row of the jacobian, Jg, is the gradient transpose $(\nabla g_i)^\top$ of g_i. In particular, if $g(x) = Mx$, then $Jg = M$. (Check this!)
(5) The dot product rule: for vector-valued functions $g(x), h(x) : \mathbb{R}^d \to \mathbb{R}^n$,

$$\nabla (g(x)^\top h(x)) = (Jg)^\top h + (Jh)^\top g$$

(6) Using these rules allows us to differentiate $f(x) = \|Mx - b\|^2 = (Mx - b) \cdot (Mx - b)$.

$$\nabla f = 2M^\top (Mx - b)$$

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

$$B^\top B \lambda = B^\top x \tag{1}$$

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

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

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

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

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

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

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

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

Remark 2.5. See examples in class or from [DFO20] of orthogonal projection matrices.

Here we see that

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

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

Example 2.6. 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

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

3. Principal components analysis

Refer to [DFO20] Chapter 10. Refer to [SSBD14], Chapter 23 for proofs.
Given $S^m = \{x_1, \dots x_m\}$ 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

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

and write

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

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.