Applications

- Data fitting and linear regression
- Least squares classification

Topics

- Orthogonal Projections and Orthogonal Subspaces (ALA 4.4)
  - Orthogonal complements
  - Orthogonality of fundamental matrix subspaces.
- Least squares
  - Problem definition (VMLS 12.1)
  - Geometric solution (LAA 6.5)
  - Computing a solution via normal equations (LAA 6.5)

Orthogonal Projections & Orthogonal Subspaces

We extend the idea of orthogonality between two vectors to orthogonality between subspaces. Our starting point is the idea of an orthogonal projection of a vector onto a subspace.

Orthogonal Projection

Let V be a (real) inner product space, and $W \subset V$ be a finite dimensional subspace of V. The results we present are fairly general, but it may be helpful to think of W as a subspace of $V = \mathbb{R}^m$.

A vector $\underline{z} \in V$ is orthogonal to the subspace $W \subset V$ if it is orthogonal to every vector in W, that is, if $\langle \underline{z}, \underline{w} \rangle = 0$ for all $\underline{w} \in W$. We will write $\underline{z} \perp W$, pronounced $\underline{z}$ "perp" W, to indicate $\underline{z}$ is perpendicular (orthogonal) to W.

A related notion is the orthogonal projection of a vector $\underline{v} \in V$ onto a subspace W, which is the element $\underline{w} \in W$ that makes the difference $\underline{z} = \underline{v} - \underline{w}$ orthogonal to W.

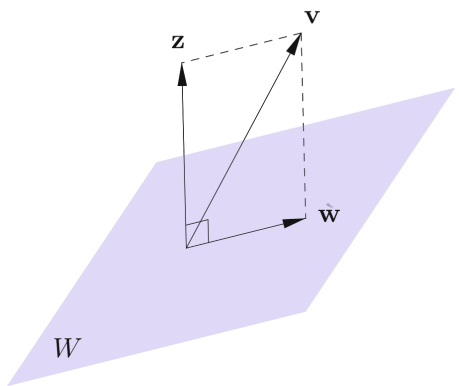

Figure 4.4. The Orthogonal Projection of a Vector onto a Subspace.

Note this means that $\underline{v}$ can be decomposed as the sum of its orthogonal projection $\underline{w} \in W$ and the perpendicular vector $\underline{z} \perp W$ that is orthogonal to W, i.e.,

$$\underline{v} = \underline{w} + \underbrace{\underline{v} - \underline{w}}_{\underline{z}} = \underline{w} + \underline{z}.$$

When we have access to an orthonormal basis for $W \subset V$, constructing the orthogonal projection of $\underline{v} \in V$ onto W becomes quite simple:

Theorem: Let $\underline{u}_1, \ldots, \underline{u}_n$ be an orthonormal basis for the subspace $W \subset V$. Then the orthogonal projection $\underline{w} \in W$ of $\underline{v} \in V$ onto W is

$$\underline{w} = c_1 \underline{u}_1 + \cdots + c_n \underline{u}_n, \quad \text{where } c_i = \langle \underline{v}, \underline{u}_i \rangle, \quad i = 1, \ldots, n$$

Proof: Since $\underline{u}_1, \dots, \underline{u}_n$ form a basis for W, we must have that $\underline{w} = c_1\underline{u}_1 + \cdots + c_n\underline{u}_n$ for some $c_1, \dots, c_n$.

If $\underline{w}$ is the orthogonal projection of $\underline{v}$ onto W, by definition we must have that $\langle \underline{v} - \underline{w}, \underline{q} \rangle = 0$ for any $\underline{q} \in W$. So let's pick $\underline{q} = \underline{u}_i$ and see what happens:

$$
\begin{aligned}
0 = \langle \underline{v} - \underline{w}, \underline{u}_i \rangle &= \langle \underline{v} - c_1\underline{u}_1 - \cdots - c_i\underline{u}_i - \cdots - c_n\underline{u}_n, \underline{u}_i \rangle \\
&= \langle \underline{v}, \underline{u}_i \rangle - c_1 \langle \underline{u}_1, \underline{u}_i \rangle - \cdots - c_i \langle \underline{u}_i, \underline{u}_i \rangle - \cdots - c_n \langle \underline{u}_n, \underline{u}_i \rangle \\
&= \langle \underline{v}, \underline{u}_i \rangle - c_i
\end{aligned}
$$

where the last line follows from $\underline{u}_1, \dots, \underline{u}_n$ being an orthonormal basis for W. Repeating for $i = 1, \dots, n$, we conclude $c_i = \langle \underline{v}, \underline{u}_i \rangle$ for $i = 1, \dots, n$ are uniquely prescribed by the orthogonality requirement, satisfying uniqueness.

Example: Consider the plane $W \subset \mathbb{R}^3$ spanned by orthogonal (but not orthonormal!) vectors

$$
\underline{v}_1 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \quad \text{and} \quad \underline{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.
$$

Let's compute the orthogonal projection of $\underline{v}$ onto $W = \operatorname{span}\{\underline{v}_1, \underline{v}_2\}$. Our first step is to normalize $\underline{v}_1$ and $\underline{v}_2$:

$$
\underline{u}_1 = \frac{\underline{v}_1}{\|\underline{v}_1\|} = \begin{bmatrix} 1/\sqrt{6} \\ -2/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix} \quad \text{and} \quad \underline{u}_2 = \frac{\underline{v}_2}{\|\underline{v}_2\|} = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}
$$

and then compute $\underline{w} = \langle \underline{v}, \underline{u}_1 \rangle \underline{u}_1 + \langle \underline{v}, \underline{u}_2 \rangle \underline{u}_2$

$$
= \frac{1}{\sqrt{6}} \begin{bmatrix} 1/\sqrt{6} \\ -2/\sqrt{6} \\ \frac{1}{\sqrt{6}} \end{bmatrix} + \frac{1}{\sqrt{3}} \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix} = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix}.
$$

We will see shortly that orthogonal projections of a vector onto a subspace is exactly what solving a least-squares problem does, and lies at the heart of machine learning and data science.

However, before that, we will explore the idea of orthogonal subspaces, and see that they provide a deep and elegant connection between the four fundamental spaces of a matrix A and whether a linear system $A\underline{x} = \underline{b}$ has a solution.

* NOTE: explain how we can write $\underline{w} = UU^T\underline{v}$ for $U = [\underline{u}_1\ \underline{u}_2 \cdots \underline{u}_n]$.

Orthogonal Subspaces

Two subspaces $W, Z \subset V$ are orthogonal if every vector in W is orthogonal to every vector in Z, that is if and only if $\langle \underline{w}, \underline{z} \rangle = 0$ for all $\underline{w} \in W$ and all $\underline{z} \in Z$.

One quick way to check this is to compare spanning sets, such as bases, for W and Z: if $W = \text{span}\{\underline{w}_1, \dots, \underline{w}_k\}$ and $Z = \text{span}\{\underline{z}_1, \dots, \underline{z}_l\}$, then W and Z are orthogonal if and only if $\langle \underline{w}_i, \underline{z}_j \rangle = 0$ for all $i = 1, \dots, k$, $j = 1, \dots, l$.

For example, if $V = \mathbb{R}^3$ and we are using the dot product, then the plane $W \subset \mathbb{R}^3$ defined by $2x - y + 3z = 0$ is orthogonal to the line Z spanned by its normal vec. $\underline{n} = (2, -1, 3)$. This is easy to check as any $\underline{w} = (x, y, z) \in W$ satisfies $\underline{n} \cdot \underline{w} = 2x - y + 3z = 0$.

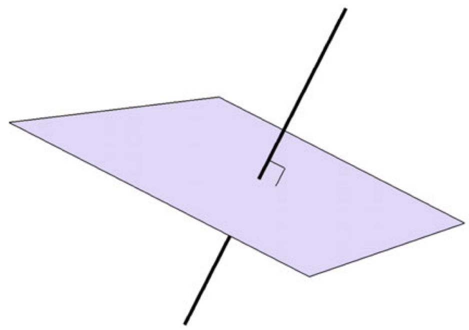

Figure 4.5. Orthogonal Complement to a Line.

An important geometric notion is the orthogonal complement $W^\perp$ of a subspace $W \subset V$, defined as the set of all vectors orthogonal to W:

$$W^\perp = \{ \underline{v} \in V \mid \langle \underline{v}, \underline{w} \rangle = 0 \text{ for all } \underline{w} \in W \}.$$

A couple of useful and easy to check properties are that:

i) $W^\perp$ is also a subspace
ii) $W \cap W^\perp = \{\underline{0}\}$, i.e., W and $W^\perp$ are transverse and only intersect at the origin.

Example: Consider again the plane $W \subset \mathbb{R}^3$ defined by the equation $2x - y + 3z = 0$. Then $W^\perp = \text{span}\{\underline{n}\} = \{(2t, -t, 3t) \mid t \in \mathbb{R}\}$ is the line spanned by its defining normal $\underline{n} = (2, -1, 3)$.

If we consider instead the set $Z = \text{span}\{\underline{n}\}$, then $Z^\perp = W$, i.e., the orthogonal complement to the line Z is the plane W. This also highlights that $Z^\perp = (W^\perp)^\perp = W$, i.e., taking the orthogonal complement twice brings you back to where you started.

Given a subspace $W \subset V$ and its orthogonal complement $W^\perp$, we can uniquely decompose any vector $\underline{v} \in V$ into $\underline{v} = \underline{w} + \underline{z}$, where $\underline{w} \in W$ and $\underline{z} \in W^\perp$. We won't prove this, but the geometric intuition is clearly conveyed in the picture below:

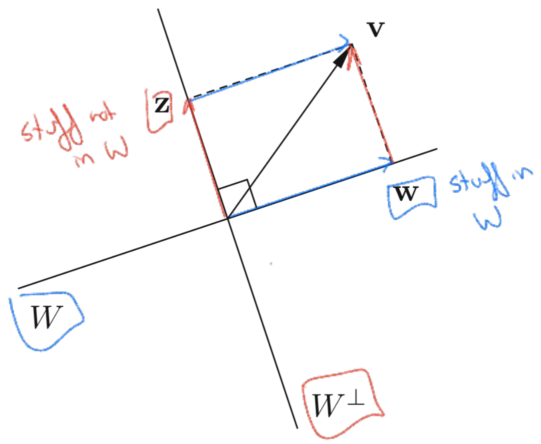

Figure 4.6. Orthogonal Decomposition of a Vector.

A useful consequence of the above, which we will use later when deriving the least squares problem solution, is that if $\underline{v} = \underline{w} + \underline{z}$, with $\underline{w} \in W$ and $\underline{z} \in W^\perp$, then $\|\underline{v}\|^2 = \|\underline{w}\|^2 + \|\underline{z}\|^2$: this is an immediate consequence of $\langle \underline{w}, \underline{z} \rangle = 0$, and is essentially Pythagoras' Theorem.

A direct consequence of this is that a subspace and its orthogonal complement have complementary dimensions:

<u>Proposition</u>: If $W \subset V$ is a subspace with $\dim W = n$ and $\dim V = m$, then $\dim W^\perp = m - n$.

If we return to our previous example where $W \subset \mathbb{R}^3$ is a plane, with $\dim W = 2$, then we can conclude that $\dim W^\perp = 1$, i.e., that $W^\perp$ is a line, which is indeed what we saw previously.

Please see online notes and ALA examples 4.42 and 4.43 for examples of decomposing a vector into elements lying in W and $W^\perp$.

<u>Orthogonality of the Fundamental Matrix Subspaces</u>

We previously introduced the four fundamental subspaces associated with an $m \times n$ matrix A, the column, null, row, and left null spaces. We also saw that the null and row spaces, are subspaces with complementary dimensions in $\mathbb{R}^n$, are the left null space and column spaces within $\mathbb{R}^m$. In fact, even more than this is true: they are orthogonal complements of each other with respect to the standard dot product.

Theorem: Let $A \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix. Then:

$$\text{Null}(A) = \text{Row}(A)^\perp = \text{Col}(A^T)^\perp \subset \mathbb{R}^n$$

and

$$\text{LNull}(A) = \text{Null}(A^T) = \text{Col}(A)^\perp \subset \mathbb{R}^m$$

We will not go through the proof (although it is not hard), but instead focus on a very important practical consequence:

Theorem: A linear system $A\underline{x} = \underline{b}$ has a solution if and only if $\underline{b}$ is orthogonal to $\text{LNull}(A)$

Ok, so what does this mean? Well remember that $A\underline{x} = \underline{b}$ if and only if $\underline{b} \in \text{Col}(A)$ since $A\underline{x}$ is a linear combination of the columns of A.

But from the above, we know that $\text{LNull}(A) = \text{Col}(A)^\perp$ or equivalently, that $\text{Col}(A) = \text{LNull}(A)^\perp = \text{Null}(A^T)^\perp$.

So this means that $\underline{b} \in \text{Null}(A^T)^\perp$, or equivalently, that $\langle \underline{y}, \underline{b} \rangle = 0$ for all $\underline{y}$ such that $A^T \underline{y} = 0$. Just to get a sense of why this is perfectly reasonable, let's assume we can find a $\underline{y} \in \text{Null}(A^T)$ for which $\langle \underline{y}, \underline{b} \rangle \neq 0$. This then immediately implies we have an inconsistent set of equations! To see this, let $\underline{x}$ be any solution to $A\underline{x} = \underline{b}$, and take the inner product of both sides with $\underline{y}$:

$$\langle \underline{y}, A\underline{x} \rangle = \langle \underline{y}, \underline{b} \rangle$$

But since $\underline{y} \in \text{LNull}(A)$, $\langle \underline{y}, A\underline{x} \rangle = \underline{y}^T A \underline{x} = 0$ for any $\underline{x}$, meaning we must have $\langle \underline{y}, \underline{b} \rangle = 0$, but we picked a special $\underline{y}$ such that $\langle \underline{y}, \underline{b} \rangle \neq 0$, so there must have been a mistake in our reasoning: either $A\underline{x} = \underline{b}$ has no solution, or $\langle \underline{y}, \underline{b} \rangle = 0$!

Another way of thinking about this: if $y^T A x = 0$, this means we can add the equations in the entries of $A\underline{x}$ together, weighted by the elements of y, so that they cancel to zero, and so the only way for $A\underline{x} = \underline{b}$ to be compatible is if the same weighted combination of the RHS, $y^T \underline{b}$, also equals 0.

Least Squares Approximation (Loosely based on VLMS 12.1 and LAA 6.5)

Suppose we are presented with an inconsistent set of linear equations $A\underline{x} = \underline{b}$. This typically coincides with $A \in \mathbb{R}^{m \times n}$ being a tall matrix, i.e., $m > n$. This corresponds to an overdetermined system of m linear equations in n unknowns. A typical setting where this arises is one of data fitting: we are given feature variables $\underline{a}_i \in \mathbb{R}^n$ and response variables $b_i \in \mathbb{R}$, and we believe that $\underline{a}_i^T \underline{x} \approx b_i$ for measurements $i = 1, \dots, m$, and $\underline{x} \in \mathbb{R}^n$ our model parameters. We will revisit this application in detail later.

The question then becomes, if no $\underline{x} \in \mathbb{R}^n$ exists such that $A\underline{x} = \underline{b}$ exists, what should we do? A natural idea is to select an $\underline{x}$ that makes the error or residual $\underline{r} = A\underline{x} - \underline{b}$ as small as possible, i.e., to find the $\underline{x}$ that minimizes $\|\underline{r}\| = \|A\underline{x} - \underline{b}\|$. Now minimizing the residual or its square gives the same answer, so we may as well minimize

$$\|A\underline{x} - \underline{b}\|^2 = \|\underline{r}\|^2 = r_1^2 + \cdots + r_m^2,$$

the sum of squares of the residuals. The problem of finding $\hat{\underline{x}} \in \mathbb{R}^n$ that minimizes $\|A\underline{x} - \underline{b}\|^2$ over all possible choices of $\underline{x} \in \mathbb{R}^n$ is called the least squares problem, and is written as

$$\text{minimize} \quad \|A\underline{x} - \underline{b}\|^2 \qquad (LS)$$

over the variable $\underline{x}$. Any $\hat{\underline{x}}$ satisfying $\|A\hat{\underline{x}} - \underline{b}\|^2 \leq \|A\underline{x} - \underline{b}\|^2$ for all $\underline{x}$ is a solution of the least-squares problem (LS), and is also called a least squares approximate solution of $A\underline{x} = \underline{b}$.

There are many ways of deriving the solution to (LS): you may have seen a vector calculus-based derivation in Math 1410. Here, we will use our new understanding of orthogonal projections to provide an intuitive and elegant geometric derivation.

Our starting point is a column interpretation of the least squares objective. Let $\underline{a}_1, \dots, \underline{a}_n \in \mathbb{R}^m$ be the columns of A: then the least squares problem (LS) is the problem of finding a linear combination of the columns that is closest to the vector $\underline{b} \in \mathbb{R}^m$, with coefficients specified by $\underline{x}$:

$$\|A\underline{x} - \underline{b}\|^2 = \|(x_1\underline{a}_1 + \cdots + x_n\underline{a}_n) - \underline{b}\|^2$$

Another way of stating this is we are seeking the vector $A\hat{\underline{x}} \in \mathrm{Col}(A)$ in the column space of A that is as close to $\underline{b}$ as possible. Perhaps not surprisingly, it turns out this can be computed by taking the orthogonal projection of $\underline{b}$ onto $\mathrm{Col}(A)$!

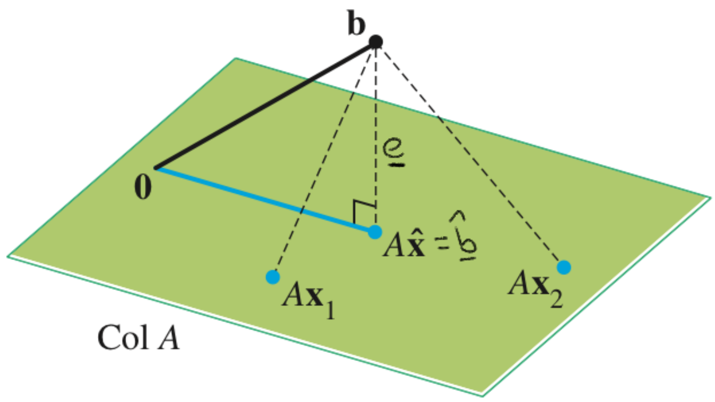

FIGURE 1 The vector **b** is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other **x**.

To prove this very geometrically intuitive fact (see Fig. 1), we need decompose $\underline{b}$ into it orthogonal projection onto Col(A), which we denote by $\hat{\underline{b}}$, and the element in its orthogonal complement Col(A), which we denote by $\underline{e}$. Recall $b, \hat{b}, \underline{e} \in \mathbb{R}^m$ and $Col(A) \subset \mathbb{R}^m$.

We then have that $\underline{r} = A\underline{x} - \underline{b} = (A\underline{x} - \hat{\underline{b}}) - (\underline{e})$. Since $A\underline{x}, \hat{\underline{b}} \in Col(A)$, so is $A\underline{x} - \hat{\underline{b}}$ (why?), and thus we have decomposed $\underline{r}$ into components lying in $Col(A)$ and $Col(A)^{\perp}$. Using our generalized Pythagorean theorem, it then follows that

$$\|A\underline{x} - \underline{b}\|^2 = \|\underline{r}\|^2 = \|A\underline{x} - \hat{\underline{b}}\|^2 + \|\underline{e}\|^2.$$

This expression can be made as small as possible by choosing $\hat{\underline{x}}$ such that $A\hat{\underline{x}} = \hat{\underline{b}}$, which always has a solution (why?) leaving the residual error $\|\underline{e}\|^2 = \|\underline{b} - \hat{\underline{b}}\|^2$, i.e., the component of $\underline{b}$ that is orthogonal to $Col(A)$.

This gives us a nice geometric interpretation of the least squares solution $\hat{\underline{x}}$, but how should we compute it?. We now recall that $Col(A)^{\perp} = Null(A^T)$, so we therefore have that $\underline{e} \in Null(A^T)$. This means that

$$A^T\underline{e} = A^T(\underline{b} - \hat{\underline{b}}) = A^T(\underline{b} - A\hat{\underline{x}}) = 0$$

or, equivalently that

$$A^TA\hat{\underline{x}} = A^T\underline{b}. \qquad (NE)$$

These are the normal equations associated with the least squares problem specified by A and $\underline{b}$. We have just informally argued that the set of least squares solutions $\hat{\underline{x}}$ coincide with the set of solutions to the normal equations (NE): this is in fact true, and can be prove (we won't do that here).

Thus we have reduced solving a least squares problem to our favorite problem, solving a system of linear equations! One question you might have is when do the normal equations (NE) have a unique solution? The answer, perhaps unsurprisingly, is when the columns of A are linearly independent, and hence form a basis for $\mathrm{Col}(A)$. The following theorem is a useful summary of our discussion thus far:

<u>Theorem</u>: Let $A \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix. Then the following statements are logically equivalent (i.e., any one being true implies all the other are true):

i) The least squares problem minimize $\|A\underline{x} - \underline{b}\|^2$ has a unique solution for any $\underline{b} \in \mathbb{R}^m$;
ii) The columns of A are linearly independent;
iii) The matrix $A^T A$ is invertible.

When these are true, the unique least squares solution is given by

$$\hat{\underline{x}} = (A^T A)^{-1} A^T \underline{b} \qquad (XLS)$$

<u>NOTE</u>: The formula (XLS) is useful mainly for theoretical purposes and for hand calculations when $A^T A$ is a 2×2 matrix. Computational approaches are typically based on QR factorizations of A (the QR factorization we saw in class for square matrices can be easily extended to tall matrices with more rows than columns).

Online notes: please include ALA example 5.12 and LAA 6.5 Examples 1-3. also add example where $A\underline{x} = \underline{b}$ has a solution and highlight that error $\underline{b} - A\hat{\underline{x}} = 0$. Ok to use np.linalg.lstsq in code examples.