Projections onto linear subspaces

Karl Stratos

Viewing vector/matrix multiplications as "projections onto linear subspaces" is one of the most useful ways to think about these operations. In this note, I'll put together necessary pieces to achieve this understanding.

1 Terminology

Given a set of linearly independent vectors $\{u_1, \ldots, u_m\} \subset \mathbb{R}^d$ where $m \leq d$, the **(linear) subspace of $\mathbb{R}^d$ spanned by** $\{u_1, \ldots, u_m\}$ is defined as:

$$\text{span}(\{u_1, \ldots, u_m\}) := \left\{ \sum_{i=1}^{m} a_i u_i \;\middle|\; a_i \in \mathbb{R} \right\}$$

In other words, $\text{span}(\{u_1, \ldots, u_m\})$ is all possible linear combinations of $u_1 \ldots u_m$. If $m = 1$, it will be an infinite line. If $m = 2$, it will be an infinite plane.

Note that for a subspace $\mathcal{V} \subset \mathbb{R}^d$ (e.g., a line, a plane, etc.), there are many choices for a set of linearly independent vectors that span $\mathcal{V}$. Any of these choices is called a **basis of $\mathcal{V}$**. The number of elements in any basis of $\mathcal{V}$ can be shown to be unique and is called the **dimension of $\mathcal{V}$**. We make two conventions (without loss of generality) that significantly simplify derivations below.

1. We always choose a basis to be orthogonal unit vectors. Such a basis is called an **orthonormal basis of $\mathcal{V}$**.
2. We always organize basis vectors into a matrix: the orthonormal basis $u_1 \ldots u_m \in \mathbb{R}^d$ is expressed as a matrix $U := [u_1 \ldots u_m] \in \mathbb{R}^{d \times m}$ whose i-th column is u_i. Note that $U^\top U = I_{m \times m}$ due to orthogonality.

Given two linearly independent vectors $u, v \in \mathbb{R}^d$, we often measure the **angle θ between u and v**. This is provided by the dot product between u and v (a result from the law of cosines):

$$u^\top v = ||u||\,||v|| \cos(\theta) \tag{1}$$

An important fact from this result is that if u and v are orthogonal, then their dot product is zero since $\cos(\pi/2) = 0$. A vector u is orthogonal to the subspace spanned by U if $u^\top v = 0$ for every $v \in \text{span}(U)$.

1.1 Projection onto a subspace

Consider some subspace of $\mathbb{R}^d$ spanned by an orthonormal basis $U = [u_1, \ldots, u_m]$. Given some $x \in \mathbb{R}^d$, a central calculation is to find $y \in \text{span}(U)$ such that $||x - y||$ is the smallest. We call this element the **projection of x onto span(U)**.

The problem can be made precise as follows. We search for a vector $y = \sum_{i=1}^{m} a_i u_i$ for some constants $a_1 \dots a_m \in \mathbb{R}$ such that the distance $||x - y||$ is minimized. The answer is given by $a_i = u_i^\top x$ and there are at least two ways to obtain this.

Method 1. Define the objective function $J(a_1 \dots a_m) = ||x - \sum_{i=1}^{m} a_i u_i||$. We can immediately see that the minimum is achieved by regression: we solve for $a = [a_1 \dots a_m]^\top \in \mathbb{R}^m$ minimizing $||x - Ua||^2$. The closed form answer is given by $a = (U^\top U)^{-1} U^\top x = U^\top x$.

Method 2. We note that the vector $x - y$ must be orthogonal to the subspace, since otherwise we can further decrease the distance. In particular, this means for each $i = 1 \dots m$ we must have: $u_i^\top (x - y) = u_i^\top \left(x - \sum_{j=1}^{m} a_j u_j \right) = u_i^\top x - a_i = 0$.

Thus the projection of x onto span(U) is given by

$$y = \sum_{i=1}^{m} (u_i^\top x) u_i = U U^\top x$$

The squared length of this projection is:

$$||y||^2 = \sum_{i=1}^{m} (u_i^\top x)^2 = \left|\left| U^\top x \right|\right|^2$$

This says that the length of $y \in \mathbb{R}^d$ is the same as the length of $U^\top x \in \mathbb{R}^m$. For this reason, the orthonormal matrix U can be viewed as defining a new coordinate system for projections. In this system, the value of the i-th coordinate is given by the length of the projection along that coordinate $u_i^\top x$:

$$U^\top x = \begin{bmatrix} u_1^\top x \\ \vdots \\ u_m^\top x \end{bmatrix}$$

To make these points clear:

- The $d \times d$ matrix (orthogonal projection) $UU^\top$ projects a point x onto span(U) in the same coordinate system, as $y = UU^\top x \in \mathbb{R}^d$.
- The $d \times m$ matrix (orthonormal basis) U projects a point x onto span(U) in a new coordinate system, as $U^\top x \in \mathbb{R}^m$.

Finally, an important property we will exploit later is that because $x - y$ is orthogonal to y by construction, the Pythagorean theorem gives us

$$||x||^2 = ||y||^2 + ||x - y||^2 \tag{2}$$

2 Choice of a subspace

At this point, we know how to project points onto a subspace spanned by an orthonormal basis U: for each x, we obtain the projection $UU^\top x$. Now given a set of points $x_1 \dots x_n \in \mathbb{R}^d$, we ask the question: out of all possible subspaces of dimension $m \leq d$, which one should we choose to project the points onto?

2.1 Best-fit subspace

One sensible choice is a subspace such that the projections onto this subspace $y_1 \dots y_n$ minimize $\sum_{l=1}^{n} ||x_l - y_l||^2$. This subspace is known as the **best-fit subspace** for $x_1 \dots x_n$.

In other words, we seek an orthonormal basis $V = [v_1 \dots v_m]$ such that:

$$V = \underset{U \in \mathbb{R}^{d \times m} \colon U^\top U = I_{m \times m}}{\arg\min} \sum_{l=1}^{n} \left|\left|x_l - UU^\top x_l\right|\right|^2$$

But since the points are constant, this is equivalent to the following by Eq. (2):

$$V = \underset{U \in \mathbb{R}^{d \times m} \colon U^\top U = I_{m \times m}}{\arg\max} \sum_{l=1}^{n} \left|\left|U^\top x_l\right|\right|^2$$

Define the "data matrix" $X := [x_1 \dots x_n]^\top \in \mathbb{R}^{n \times d}$ whose l-th row is $x_l^\top$. Then $\sum_{l=1}^{n} \left|\left|U^\top x_l\right|\right|^2 = ||XU||_F^2 = \mathrm{Tr}(U^\top X^\top XU)$, so the above problem can be framed as:

$$V = \underset{U \in \mathbb{R}^{d \times m} \colon U^\top U = I_{m \times m}}{\arg\max} \mathrm{Tr}(U^\top X^\top XU)$$

This solution is given by the singular value decomposition (SVD) of $X = \bar{U}\bar{\Sigma}\bar{V}^\top$ where $\bar{\Sigma}$ is the diagonal matrix of singular values in decreasing magnitude. Specifically, V is the first m columns of $\bar{V}$, i.e., the right singular vectors of X corresponding to the largest m singular values. See Appendix 3 for a discussion. Consequently, the projections onto the best-fit subspace are given by the rows of

$$XV = \bar{U}_m \bar{\Sigma}_m$$

where $\bar{U}_m$ is the first m columns of $\bar{U}$ and $\bar{\Sigma}_m$ is the first $m \times m$ block of $\bar{\Sigma}$ in the upper left corner.

2.1.1 Relation to principal component analysis (PCA)

In general, the best-fit subspace may be useless in characterizing arbitrary $x_1 \dots x_n$. This is because a subspace must pass through the origin. To see how this can be problematic, suppose we have three points $(-1, 1)$, $(0, 1)$, and $(1, 1)$. Even though they are in $\mathbb{R}^2$, it is obvious they are essentially points $-1, 0, 1$ in $\mathbb{R}$. But the best-fit subspace is given by the y-axis. Consequently, all points collapse to $(0, 1)$ when projected to the best-fit subspace.

A remedy is to first preprocess points so that they are centered at the origin: we subtract $(1/n)\sum_{l=1}^{n} x_l$ from each x_l. The above example would then be $(-1, 0)$, $(0, 0)$, and $(1, 0)$ and the best-fit subspace is now given by the x-axis.

Finding the best-fit subspace of "centered" points is called PCA. It has an interpretation of maximizing the variance of projections along each orthogonal axis.

2.1.2 Relation to least squares

Least squares refers to the minimization of squared loss between certain input and target variables. The original dimension is partitioned as $d = m_1 + m_2$ and then m_1

input coordinates and m_2 target coordinates are selected. Thus a point $x_l \in \mathbb{R}^d$ is divided into an input part $\bar{x}_l \in \mathbb{R}^{m_1}$ and a target part $y_l \in \mathbb{R}^{m_2}$. Then we compute

$$\arg\min_{U \in \mathbb{R}^{m_1 \times m_2}} \sum_{l=1}^{n} \left|\left|y_l - U^\top \bar{x}_l\right|\right|^2$$

This objective is more compactly represented as $\left|\left|Y - \bar{X}U\right|\right|_F^2$ where $Y := [y_1 \dots y_n]^\top$ and $\bar{X} := [\bar{x}_1 \dots \bar{x}_n]^\top$ and the solution is given by $U = (X^\top X)^{-1} X^\top Y$.

While least squares is similar to finding the best-fit subspace in the sense that they both minimize the sum of squared differences, they are clearly different in major ways. First, the best-fit subspace is selected based on all coordinates: in contrast, in least squares, coordinates must be partitioned to input and target variables. Second, there is no orthonormal constraint on the parameter U in least squares. Third, finding the best-fit subspace is equivalent to the problem of low-rank approximation which is a non-convex problem (despite this non-convexity, SVD can find an exact solution): on the other hand, least squares is a convex problem.

2.2 Random subspace

Perhaps surprisingly, another sensible choice of subspace is a *random* subspace chosen independently of the data. This might not sound as sensible as the best-fit subspace since it does not consider the information in the data. But a crucial advantage of this approach is that due to its randomness, adversaries are prevented from hindering our objective by manipulating the data.

An important result is the following.

Theorem 2.1 (Johnson-Lindenstrauss lemma)**.** *Let $x_1 \dots x_n \in \mathbb{R}^d$. Then there is a distribution $\mathcal{D}$ over $\mathbb{R}^{d \times m}$ where $m = O(\log n)$ such that if $U \sim \mathcal{D}$, then*

$$\left|\left|U^\top x_i - U^\top x_j\right|\right|^2 \approx ||x_i - x_j||^2$$

for all $i \neq j$ with high probability.

That is, we can reduce the dimension of data points $x_1 \dots x_n$ from d to $O(\log n)$ while preserving pairwise distances $||x_i - x_j||^2$ (with high probability). This is accomplished by linearly embedding the points onto a random m-dimensional subspace.[1] Furthermore, the reduced dimension m does not depend on the original dimension d at all, only the number of data points n. Thus if d is as large as n, this yields an exponential reduction in dimension.

When should we prefer random subspaces over PCA? Dasgupta (2000) gives an argument based on clustering under a mixture of $2d$ Gaussians in $\mathbb{R}^d$. It is easy to adversarially configure the Gaussians (e.g., symmetrically along each axis) so that PCA must collapse some of them together in any lower-dimensional subspace, thereby destroying any hope of separating the collapsed Gaussians. In contrast, a random projection is guaranteed to preserve pairwise separations independently of d. This is an example of how randomness prevents adversarial configurations of data.

[1] Since $U \in \mathbb{R}^{d \times m}$ is not required to be orthonormal, $UU^\top$ is generally not a projection in $\mathbb{R}^d$ unlike the best-fit projections. But clearly $U^\top x \in \mathbb{R}^m$ resides in some m-dimensional subspace of $\mathbb{R}^d$, so we will just say U "linearly embeds" x onto it.

2.2.1 A formulation of the Johnson-Lindenstrauss lemma

Here is a more formal statement of Theorem 2.1.

Theorem 2.2. *Given $0 < \epsilon < 1$ and $0 < \delta < 1$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^{d \times m}$ where $m = O(\epsilon^{-2} \log(n/\delta))$ such that if $U \sim \mathcal{D}$, then*

$$(1-\epsilon)\left|\left|x_i - x_j\right|\right|^2 \leq \left|\left|U^\top x_i - U^\top x_j\right|\right|^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|^2$$

for all $i \neq j$ with probability at least $1 - \delta$.

It turns out that there are many possible distributions $\mathcal{D}$ that can be used to prove Theorem 2.2. One such formulation is the following (a sketch proof is provided in Appendix 4).

Theorem 2.3 (Indyk and Motwani (1998))**.** *Let $0 < \epsilon < 1$ and $0 < \delta < 1$. Pick $U \in \mathbb{R}^{d \times m}$ where each term is randomly drawn as $U_{i,j} \sim \mathcal{N}(0, 1/m)$ for $m = O(\epsilon^{-2} \log(n/\delta))$. Then*

$$(1-\epsilon)\left|\left|x_i - x_j\right|\right|^2 \leq \left|\left|U^\top x_i - U^\top x_j\right|\right|^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|^2$$

for all $i \neq j$ with probability at least $1 - \delta$.

3 Appendix: SVD for optimization

A useful characterization of SVD is the maximization of the norm or the trace. In particular, we can solve the following problem: given $X \in \mathbb{R}^{n \times d}$, find a matrix $V = [v_1 \dots v_m] \in \mathbb{R}^{d \times m}$ such that

$$V = \underset{U \in \mathbb{R}^{d \times m}:\ U^\top U = I_{m \times m}}{\arg\max} \left|\left|XU\right|\right|_F^2$$

Since $\left|\left|M\right|\right|_F = \sqrt{\mathrm{Tr}(M^\top M)}$, this is the same as

$$V = \underset{U \in \mathbb{R}^{d \times m}:\ U^\top U = I_{m \times m}}{\arg\max} \mathrm{Tr}(U^\top X^\top X U) \tag{3}$$

Let $X = \bar{U}\bar{\Sigma}\bar{V}^\top$ be the standard SVD of X (assume $n \geq d$ wlog):

$$\bar{U} = [\bar{u}_1 \dots \bar{u}_d] \in \mathbb{R}^{n \times d} \qquad \bar{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_d) \in \mathbb{R}^{d \times d} \qquad \bar{V} = [\bar{v}_1 \dots \bar{v}_d] \in \mathbb{R}^{d \times d}$$

where $\sigma_1 \geq \dots \geq \sigma_d \geq 0$ and $\bar{U}$ and $\bar{V}$ have orthonormal columns. Then $U^\top X^\top X U = U^\top \bar{V}\bar{\Sigma}^2\bar{V}^\top U$ and (3) can be expressed in terms of the right singular vectors:

$$v_1 \dots v_m = \underset{\substack{u_1 \dots u_m \in \mathbb{R}^d: \\ u_i^\top u_i = 1 \ \ \forall i \\ u_i^\top u_j = 0 \ \ \forall i \neq j}}{\arg\max} \sum_{i=1}^{m} \sum_{l=1}^{d} \sigma_l^2 (u_i^\top \bar{v}_l)^2 \tag{4}$$

We will need the following lemma.

Lemma 3.1. *Let θ be an angle in $[0, \pi/2]$. Let $a \geq b$ be constants. Then*

$$a\cos^2(\theta) + b\sin^2(\theta)$$

is maximized at $\theta = 0$.

Proof. The objective is concave in θ. Setting the derivative with respect to θ to zero, we arrive in the equation

$$a\sin(2\theta) = b\sin(2\theta)$$

implying that $\sin(2\theta) = 0$ and thus $\theta = 0$. Note that if $a = b$ then θ can be anything (including 0). □

Theorem 3.2. *For all $1 \le m \le d$, a solution of (4) is given by $v_i = \bar{v}_i$ for $i = 1 \dots m$.*

Proof. When $m = 1$, we seek $v_1 \in \mathbb{R}^d$ such that

$$v_1 = \operatorname*{arg\,max}_{\substack{u \in \mathbb{R}^d: \\ u^\top u = 1}} \sum_{l=1}^{d} \sigma_l^2 (u^\top \bar{v}_l)^2 \tag{5}$$

Let θ_l denote the angle between u and $\bar{v}_l$. Since both u and $\bar{v}_l$ are unit vectors, we have $\cos(\theta_l) = u^\top \bar{v}_l$ by (1). Now select any $a \neq b$ in $\{1, \dots, d\}$ where $\sigma_a \ge \sigma_b$. Since $\bar{v}_a$ and $\bar{v}_b$ are orthogonal, we have $\theta_b = \pi/2 - \theta_a$ and $\cos(\theta_b) = \sin(\theta_a)$. At maximum of (5), we must have

$$\theta_a = \operatorname*{arg\,max}_{\theta \in [0, \pi/2]} \sigma_a^2 \cos^2(\theta) + \sigma_b^2 \sin^2(\theta)$$

By Lemma 3.1, $\theta_a = 0$ and $\theta_b = \pi/2$. Apply this between $a = 1$ and $b = 1 \dots d$ and conclude that $\theta_1 = 0$ and $\theta_l = \pi/2$ for all $l \neq 1$. Thus $v_1 = \bar{v}_1$.

Assume the statement is true for some $m = C$ where $1 \le C < d$. We consider the case $m = C + 1$. For convenience, define

$$S = \{u_1 \dots u_{C+1} \in \mathbb{R}^d : \ u_i^\top u_i = 1 \ \ \forall i, \ \ u_i^\top u_j = 0 \ \ \forall i \neq j\}$$
$$S' = \{u_1 \dots u_C \quad \in \mathbb{R}^d : \ u_i^\top u_i = 1 \ \ \forall i, \ \ u_i^\top u_j = 0 \ \ \forall i \neq j\}$$

Observe that the maximum of (4) is achieved at:

$$\max_{u_1 \dots u_{C+1} \in S} \ \sum_{i=1}^{C} \sum_{l=1}^{d} \sigma_l^2 (u_i^\top \bar{v}_l)^2 + \sum_{l=1}^{d} \sigma_l^2 (u_{C+1}^\top \bar{v}_l)^2$$
$$= \max_{u_1 \dots u_C \in S'} \ \sum_{i=1}^{C} \sum_{l=1}^{d} \sigma_l^2 (u_i^\top \bar{v}_l)^2 + \max_{\substack{u \in \mathbb{R}^d: \\ u^\top u = 1 \\ u^\top u_i = 0 \ \ \forall i = 1 \dots C}} \ \sum_{l=1}^{d} \sigma_l^2 (u^\top \bar{v}_l)^2$$

The two terms can be treated independently. The first term is maximized at $\bar{v}_1 \dots \bar{v}_m$ by assumption. Thus the solution is given by $v_i = \bar{v}_i$ for $i = 1 \dots C$ and

$$v_{C+1} = \operatorname*{arg\,max}_{\substack{u \in \mathbb{R}^d: \\ u^\top u = 1 \\ u^\top u_i = 0 \ \ \forall i = 1 \dots C}} \ \sum_{l=1}^{d} \sigma_l^2 (u^\top \bar{v}_l)^2$$

It is easy to verify that $v_{C+1} = \bar{v}_{C+1}$ using an argument similar to the above. □

4 Appendix: proof of Theorem 2.3

The proof of Theorem 2.3 relies on the following lemma.

Lemma 4.1. *Let $z \in \mathbb{R}^d$ be any unit vector. There exists a constant C such that given $0 < \epsilon < 1$ and $0 < \delta < 1$, if we pick $m \geq (C/\epsilon^2)\log(1/\delta)$ then the matrix $U \in \mathbb{R}^{d\times m}$ defined as in Theorem 2.3 satisfies*

$$\left| \left|\left|U^\top z\right|\right|^2 - 1 \right| > \epsilon$$

with probability at most δ.

In other words, U preserves the norm of unit vectors. Using this lemma, it is straightforward to prove the main theorem, which is re-stated here for convenience.

Theorem (Indyk and Motwani (1998))**.** *Let $0 < \epsilon < 1$ and $0 < \delta < 1$. Pick $U \in \mathbb{R}^{d\times m}$ where each term is randomly drawn as $U_{i,j} \sim \mathcal{N}(0, 1/m)$ for $m = O(\epsilon^{-2}\log(n/\delta))$. Then*

$$(1-\epsilon)\left|\left|x_i - x_j\right|\right|^2 \leq \left|\left|U^\top x_i - U^\top x_j\right|\right|^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|^2$$

for all $i \neq j$ with probability at least $1 - \delta$.

Proof. By Lemma 4.1, there is some constant C such that for particular $i \neq j$, we can apply Lemma 4.1 to the unit vector $(x_i - x_j)/\left|\left|x_i - x_j\right|\right|$ to obtain

$$\left|\left|U^\top x_i - U^\top x_j\right|\right|^2 > (1+\epsilon)\left|\left|x_i - x_j\right|\right|^2 \qquad \text{or}$$
$$\left|\left|U^\top x_i - U^\top x_j\right|\right|^2 < (1-\epsilon)\left|\left|x_i - x_j\right|\right|^2$$

with probability at most δ' where $m \geq (C/\epsilon^2)\log(1/\delta')$. By the union bound, the probability that this happens for some $i \neq j$ is at most $\binom{n}{2}\delta'$. Thus for all $i \neq j$

$$(1-\epsilon)\left|\left|x_i - x_j\right|\right|^2 \leq \left|\left|U^\top x_i - U^\top x_j\right|\right|^2 \leq (1+\epsilon)\left|\left|x_i - x_j\right|\right|^2$$

with probability at least $1 - \binom{n}{2}\delta'$. Solving for δ' in $m \geq (C/\epsilon^2)\log(1/\delta')$, we obtain $\delta' \geq \exp(-C^{-1}m\epsilon^2)$. Thus

$$\delta = \binom{n}{2}\delta' \geq \frac{n^2-n}{2}\exp\left(-\frac{1}{C}m\epsilon^2\right)$$

When we solve for m, we obtain

$$m \geq \frac{C}{\epsilon^2}\log\left(\frac{n^2-n}{2\delta}\right)$$

Thus the statement holds for $m = O(\epsilon^{-2}\log(n/\delta))$. □

Sketch proof of Lemma 4.1. Define $X = \sqrt{m}U^\top z$ (which is a random variable because U is random). Note that $X_i = \sqrt{m}z^\top u_i$ for independent $u_i \sim \mathcal{N}(0, m^{-1}I_{m\times m})$ with

$$\mathbf{E}[X_i] = 0 \qquad \mathbf{E}[X_i^2] = 1$$

Thus each X_i is distributed as $\mathcal{N}(0,1)$ and $\left|\left|X\right|\right|^2 = m\left|\left|U^\top z\right|\right|^2$ is distributed as $\chi^2(d)$. A Chernoff bound on $\chi^2(d)$ gives the desired result. □