Singular value decomposition

COMS 3251 Fall 2022 (Daniel Hsu)

1 Best fitting subspaces

1.1 Problem definition

A typical data science problem represents data points as tuples of numerical attribute values. Suppose a data set is composed of m data points, $\mathbf{a}_1, \ldots, \mathbf{a}_m$, with each $\mathbf{a}_i$ being an n-tuple of real numbers, so n denotes the number of attributes per data point. In cases where n is very large, it is potentially helpful to reduce the number of attributes to something much smaller. Of course, this cannot be done indiscriminately; we would want the "reduced" data points to faithfully represent the original data points in some manner.

If the data points are regarded as vectors in n-dimensional Euclidean space $\mathbb{R}^n$, a natural approach to "reduce" each data point is to project them to a k-dimensional subspace W of $\mathbb{R}^n$, for some $k \leq n$ (and ideally, $k \ll n$). If $\mathcal{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_k)$ is an ordered ONB for W, then for any n-vector $\mathbf{v}$, the coordinate representation for the orthogonal projection of $\mathbf{v}$ to W with respect to $\mathcal{Q}$ is the k-vector $(\langle \mathbf{q}_1, \mathbf{v} \rangle, \ldots, \langle \mathbf{q}_k, \mathbf{v} \rangle)$. Using this coordinate representation, we can reconstruct the vector in W that is closest to $\mathbf{v}$ in Euclidean distance:

$$P_W \mathbf{v} \;=\; \langle \mathbf{q}_1, \mathbf{v} \rangle \, \mathbf{q}_1 + \cdots + \langle \mathbf{q}_k, \mathbf{v} \rangle \, \mathbf{q}_k.$$

Above, P_W denotes the orthoprojector for the subspace W. See Figure 1.

So, we may seek to find the k-dimensional subspace W such that the Euclidean distances between the original data points and their orthogonal projections to W are as small as possible. There are m such distances, and one way of putting all of them together into a single quality measure is to consider their sum of squares:

$$\mathrm{cost}(W; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;=\; \|\mathbf{a}_1 - P_W \mathbf{a}_1\|^2 + \cdots + \|\mathbf{a}_m - P_W \mathbf{a}_m\|^2.$$

The cost is non-negative, and is zero if and only if $\mathbf{a}_i \in W$ for all $i \in \{1, \ldots, m\}$. We call the problem of finding a k-dimensional subspace of minimum cost the *best fitting k-dimensional subspace problem (k-BFS)*.

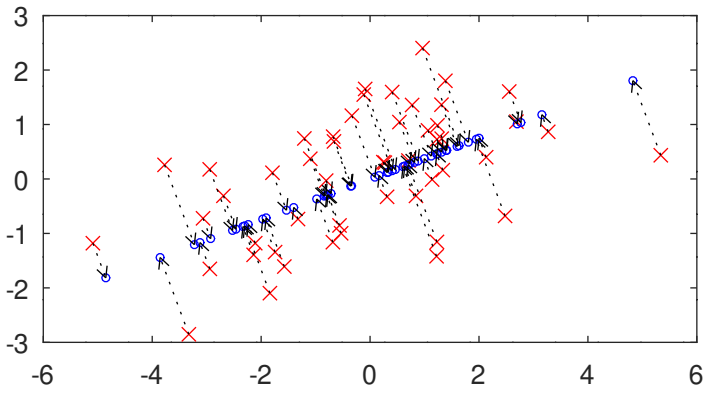

Figure 1: Data set of $m = 50$ data points, each represented as a vector in n-dimensional Euclidean space for $n = 2$. The data points are shown as red $\times$'s, and each is approximated by its orthogonal projection to the best fitting k-dimensional subspace for $k = 1$; the projections are shown as blue $\circ$'s.

1.2 Reformulation of the cost

Recall, that for any subspace W of $\mathbb{R}^n$, the orthoprojector P_W decomposes any vector $\mathbf{v} \in \mathbb{R}^n$ uniquely into a part in W and a part orthogonal to W:

$$\mathbf{v} \;=\; P_W\mathbf{v} + (I - P_W)\mathbf{v}.$$

By the Pythagorean Theorem,

$$\|\mathbf{v}\|^2 \;=\; \|P_W\mathbf{v}\|^2 + \|(I - P_W)\mathbf{v}\|^2 \;=\; \|P_W\mathbf{v}\|^2 + \|\mathbf{v} - P_W\mathbf{v}\|^2.$$

Therefore, for a given data set $\mathbf{a}_1, \ldots, \mathbf{a}_m$, the cost of W is

$$\begin{aligned}
\mathrm{cost}(W; \mathbf{a}_1, \ldots, \mathbf{a}_m) &= \|\mathbf{a}_1 - P_W\mathbf{a}_1\|^2 + \cdots + \|\mathbf{a}_m - P_W\mathbf{a}_m\|^2 \\
&= \left(\|\mathbf{a}_1\|^2 - \|P_W\mathbf{a}_1\|^2\right) + \cdots + \left(\|\mathbf{a}_m\|^2 - \|P_W\mathbf{a}_m\|^2\right) \\
&= \left(\|\mathbf{a}_1\|^2 + \cdots + \|\mathbf{a}_m\|^2\right) - \left(\|P_W\mathbf{a}_1\|^2 + \cdots + \|P_W\mathbf{a}_m\|^2\right).
\end{aligned}$$

The first parenthesized term on the final right-hand side isn't affected by W. So minimizing the cost is the same as maximizing the gain defined by

$$\mathrm{gain}(W; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;=\; \|P_W\mathbf{a}_1\|^2 + \cdots + \|P_W\mathbf{a}_m\|^2.$$

There is another useful way to understand the gain. Identify a k-dimensional subspace with an ONB, say, $\mathbf{q}_1, \ldots, \mathbf{q}_k$, and adopt the shorthand

$$\mathrm{gain}(\mathbf{q}_1, \ldots, \mathbf{q}_k; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;=\; \mathrm{gain}(\mathrm{span}(\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}); \mathbf{a}_1, \ldots, \mathbf{a}_m).$$

By Parseval's identity, for any $\mathbf{v} \in \mathbb{R}^n$,

$$\|P_W \mathbf{v}\|^2 \;=\; \langle \mathbf{v}, \mathbf{q}_1 \rangle^2 + \cdots + \langle \mathbf{v}, \mathbf{q}_k \rangle^2.$$

So, using a double summation,

$$\begin{aligned} \mathrm{gain}(\mathbf{q}_1, \ldots, \mathbf{q}_k; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;&=\; \sum_{i=1}^{m} \sum_{j=1}^{k} \langle \mathbf{a}_i, \mathbf{q}_j \rangle^2 \;=\; \sum_{j=1}^{k} \sum_{i=1}^{m} \langle \mathbf{a}_i, \mathbf{q}_j \rangle^2 \\ &=\; \sum_{j=1}^{k} \mathrm{gain}(\mathbf{q}_j; \mathbf{a}_1, \ldots, \mathbf{a}_m) \end{aligned}$$

where $\mathrm{gain}(\mathbf{x}; \mathbf{a}_1, \ldots, \mathbf{a}_m) = \sum_{i=1}^{m} \langle \mathbf{a}_i, \mathbf{x} \rangle^2$. The decomposition of the gain into the k terms will be useful in our analysis.[1]

1.3 Greedy algorithm and best fitting lines

We claim that the following algorithm solves k-BFS.

Algorithm 1 Greedy algorithm for best fitting k-dimensional subspace

Input: Data points $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$.

1: **for** $j = 1, \ldots, k$ **do**
2: Let $S_{j-1} = \mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_{j-1}\})$.
3: Let $\mathbf{v}_j$ maximize $\mathrm{gain}(\mathbf{x}; \mathbf{a}_1, \ldots, \mathbf{a}_m)$ among all unit vectors $\mathbf{x} \in S_{j-1}^{\perp}$.
4: **end for**
5: **return** $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

In each iteration of Algorithm 1 (specifically, in Line 3), a sub-problem resembling the $k = 1$ version of k-BFS must be solved. It is not exactly the same as 1-BFS on account of the constraint $\mathbf{x} \in S_{j-1}^{\perp}$.

Example. Consider the data set $\mathbf{a}_1 = (3, 1, 2, 0)$, $\mathbf{a}_2 = (-1, -3, 0, -2)$, $\mathbf{a}_3 = (0, 2, 1, 3)$, and $\mathbf{a}_4 = (-2, 0, -3, -1)$. Suppose we seek a 2-dimenisional subspace to approximately fit the data points.

[1] We may drop the dependence of $\mathrm{cost}(\,\cdot\,; \mathbf{a}_1, \ldots, \mathbf{a}_m)$ and $\mathrm{gain}(\,\cdot\,; \mathbf{a}_1, \ldots, \mathbf{a}_m)$ on the data set $\mathbf{a}_1, \ldots, \mathbf{a}_m$ when it is clear from context, and simply write $\mathrm{cost}(\,\cdot\,)$ and $\mathrm{gain}(\,\cdot\,)$.

- Iteration $j = 1$:

$$\begin{aligned} S_0 &= \{\mathbf{0}\}, \\ S_0^\perp &= \mathbb{R}^4, \\ \mathbf{v}_1 &= (1/2, 1/2, 1/2, 1/2). \end{aligned}$$

  This achieves $\text{gain}(\mathbf{v}_1) = 3^2 + (-3)^2 + 3^2 + (-3)^2 = 36$.

- Iteration $j = 2$:

$$\begin{aligned} S_1 &= \{(c, c, c, c) : c \in \mathbb{R}\}, \\ S_1^\perp &= \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : x_1 + x_2 + x_3 + x_4 = 0\}, \\ \mathbf{v}_2 &= (1/2, -1/2, 1/2, -1/2). \end{aligned}$$

  This achieves $\text{gain}(\mathbf{v}_2) = 2^2 + 2^2 + (-2)^2 + (-2)^2 = 16$.

Note that $\|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 + \|\mathbf{a}_3\|^2 + \|\mathbf{a}_4\|^2 = 56$. This means that the final cost is $56 - 36 - 16 = 4$. ∎

If we have a subroutine for solving 1-BFS—the *best fitting line problem (BFL)*—we can solve the required sub-problem by suitably modifying the data set. The sub-problem is to maximize $\text{gain}(\mathbf{x}; \mathbf{a}_1, \ldots, \mathbf{a}_m)$ among all unit vectors orthogonal to (the span of) the previous chosen unit vectors. If P_{j-1} denotes the orthoprojector for S_{j-1}, then each $\mathbf{a}_i$ can be written as

$$\mathbf{a}_i = P_{j-1}\mathbf{a}_i + (I - P_{j-1})\mathbf{a}_i,$$

and hence if $\mathbf{x} \in S_{j-1}^\perp$, then

$$\langle \mathbf{a}_i, \mathbf{x} \rangle = \langle P_{j-1}\mathbf{a}_i + (I - P_{j-1})\mathbf{a}_i, \mathbf{x} \rangle = \cancelto{0}{\langle P_{j-1}\mathbf{a}_i, \mathbf{x} \rangle} + \langle (I - P_{j-1})\mathbf{a}_i, \mathbf{x} \rangle.$$

This means the gain of $\mathbf{x} \in S_{j-1}^\perp$ satisfies

$$\begin{aligned} \text{gain}(\mathbf{x}; \mathbf{a}_1, \ldots, \mathbf{a}_m) &= \sum_{i=1}^m \langle \mathbf{a}_i, \mathbf{x} \rangle^2 = \sum_{i=1}^m \langle (I - P_{j-1})\mathbf{a}_i, \mathbf{x} \rangle^2 \\ &= \text{gain}(\mathbf{x}; \mathbf{b}_1, \ldots, \mathbf{b}_m), \end{aligned}$$

where $\mathbf{b}_i = (I - P_{j-1})\mathbf{a}_i$. The final expression is the gain of $\mathbf{x}$ on a modified data set obtained by projecting each $\mathbf{a}_i$ to the orthogonal complement of S_{j-1}.

On the other hand, even if the unit vector $\mathbf{x}$ was not restricted to be in S_{j-1}, the gain achieved on the modified data set $\mathbf{b}_1, \ldots, \mathbf{b}_m$ is

$$\begin{aligned}
\text{gain}(\mathbf{x}; \mathbf{b}_1, \ldots, \mathbf{b}_m) &= \sum_{i=1}^{m} \langle (I - P_{j-1})\mathbf{a}_i, \mathbf{x} \rangle^2 \\
&= \sum_{i=1}^{m} \langle (I - P_{j-1})^2 \mathbf{a}_i, \mathbf{x} \rangle^2 \\
&= \sum_{i=1}^{m} \langle (I - P_{j-1})\mathbf{a}_i, (I - P_{j-1})\mathbf{x} \rangle^2 \\
&= \text{gain}((I - P_{j-1})\mathbf{x}; \mathbf{b}_1, \ldots, \mathbf{b}_m).
\end{aligned}$$

Above, the second equality uses the idempotency property of the projector $I - P_{j-1}$. So, it is not possible to achieve any higher gain value by allowing the unit vector $\mathbf{x}$ to have a non-zero component in S_{j-1}, and hence, it suffices to maximize the gain on the modified data set over all unit vectors $\mathbf{x}$.

1.4 Optimality of the greedy algorithm

Let us not worry about the difficulty of solving the BFL problem for now (see Appendix A), and instead let us forge ahead with analyzing Algorithm 1.

The following theorem is main performance guarantee for Algorithm 1.

Theorem 1. *The execution of Algorithm 1 on a data set $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ returns orthonormal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in $\mathbb{R}^n$ such that their span S_k satisfies*

$$\text{gain}(S_k; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;\geq\; \text{gain}(W; \mathbf{a}_1, \ldots, \mathbf{a}_m)$$

for all k-dimensional subspaces W of $\mathbb{R}^n$. Moreover, the non-negative values

$$\sigma_j \;=\; \sqrt{\text{gain}(\mathbf{v}_j; \mathbf{a}_1, \ldots, \mathbf{a}_m)} \quad \text{for } j \in \{1, \ldots, k\}$$

satisfy

$$\sigma_1 \;\geq\; \cdots \;\geq\; \sigma_k.$$

Proof. We first prove the optimality property of S_k. The proof is by induction on k. The claim is true for $k = 1$ by definition of $\mathbf{v}_1$ in Line 3.

So, for some $k \geq 2$, assume as the inductive hypothesis that

$$\text{gain}(S_{k-1}) \;\geq\; \text{gain}(W')$$

for all $(k-1)$-dimensional subspaces W' of $\mathbb{R}^n$. Consider any k-dimensional subspace W of $\mathbb{R}^n$. We need to show that

$$\text{gain}(S_k) \;\geq\; \text{gain}(W).$$

If $W \cap S_{k-1}^{\perp} = \{\mathbf{0}\}$, then $\dim(W) \leq n - \dim(S_{k-1}^{\perp}) = k-1$, a contradiction of the assumed dimensionality of W. Hence, we may assume there exists a non-zero vector $\mathbf{w} \in W \cap S_{k-1}^{\perp}$; let $\mathbf{q}_1 = \mathbf{w}/\|\mathbf{w}\|$. By the ONB Completion Theorem, there exist unit vectors $\mathbf{q}_2, \ldots, \mathbf{q}_k \in W$ such that $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ is an ONB for W. Let $W^- = \text{span}(\{\mathbf{q}_2, \ldots, \mathbf{q}_k\})$, and observe that

$$\text{gain}(W) \;=\; \text{gain}(\mathbf{q}_1) + \sum_{j=2}^{k} \text{gain}(\mathbf{q}_j) \;=\; \text{gain}(\mathbf{q}_1) + \text{gain}(W^-).$$

Since $\mathbf{q}_1 \in S_{k-1}^{\perp}$, the choice of $\mathbf{v}_k$ in Line 3 implies

$$\text{gain}(\mathbf{q}_1) \;\leq\; \text{gain}(\mathbf{v}_k).$$

And since $\dim(W^-) = k-1$, the inductive hypothesis implies

$$\text{gain}(W^-) \;\leq\; \text{gain}(S_{k-1}).$$

We conclude that

$$\begin{aligned}
\text{gain}(W) \;&=\; \text{gain}(\mathbf{q}_1) + \text{gain}(W^-) \\
&\leq\; \text{gain}(\mathbf{v}_k) + \text{gain}(S_{k-1}) \\
&=\; \text{gain}(\mathbf{v}_k) + \sum_{j=1}^{k-1} \text{gain}(\mathbf{v}_j) \;=\; \text{gain}(S_k).
\end{aligned}$$

So, by the principle of mathematical induction, the first part of the theorem is proven.

For the second part, we assume for sake of contradiction that for some pair of indices (i, j) with $1 \leq i < j \leq k$, we have $\sigma_i < \sigma_j$. Take the pair

(i, j) for which i is as small as possible. Since $S_{i-1} \subseteq S_{j-1}$, it follows that $S_{j-1}^{\perp} \subseteq S_{i-1}^{\perp}$, so $\mathbf{v}_j \in S_{j-1}^{\perp} \subseteq S_{i-1}^{\perp}$. On the other hand, $\mathbf{v}_i$ has the highest gain among unit vectors in $S_{i-1}^{\perp}$, as per Line 3 in Algorithm 1. Therefore

$$\text{gain}(\mathbf{v}_i) \;\geq\; \text{gain}(\mathbf{v}_j).$$

But this contradicts the inequality $\sigma_i < \sigma_j$. Hence we must conclude that no such pair (i, j) exists. This proves the second part of the theorem. □

1.5 Additional properties of Algorithm 1

In this section, we prove additional properties of the ONB returned by Algorithm 1. In the following, we consider an arbitrary data set $\mathbf{a}_1, \ldots, \mathbf{a}_m$ from $\mathbb{R}^n$, and define

$$r \;=\; \dim(\text{span}(\{\mathbf{a}_1, \ldots, \mathbf{a}_m\})).$$

We consider the execution of Algorithm 1 on this data set, with $k = r$.

Proposition 1. *Consider the setting of Theorem 1 and Section 1.5. Then*

1. $S_r = \text{span}(\{\mathbf{a}_1, \ldots, \mathbf{a}_m\})$;
2. $\text{cost}(S_k) = \sum_{j=k+1}^{r} \sigma_j^2$ *for any* $k \in \{0, \ldots, r\}$.
3. $\sigma_j > 0$ *for all* $j \in \{1, \ldots, r\}$.

Proof. By Theorem 1,

$$\text{cost}(S_r) \;\leq\; \text{cost}(\text{span}(\{\mathbf{a}_1, \ldots, \mathbf{a}_m\})) \;=\; 0.$$

Hence it must be that $\text{cost}(S_r) = 0$, and $S_r = \text{span}(\{\mathbf{a}_1, \ldots, \mathbf{a}_m\})$, proving the first claim. Note that

$$\sigma_j^2 \;=\; \text{gain}(\mathbf{v}_j) \;=\; \text{cost}(S_{j-1}) - \text{cost}(S_j). \tag{1}$$

For any $k \in \{0, \ldots, r\}$, summing (1) over $j \in \{k+1, \ldots, r\}$ gives

$$\sum_{j=k+1}^{r} \sigma_j^2 \;=\; \text{cost}(S_k) - \text{cost}(S_r).$$

Since $\text{cost}(S_r) = 0$, it follows that $\text{cost}(S_k) = \sum_{j=k+1}^{r} \sigma_j^2$. This proves the second claim.

Suppose for sake of contradiction that $\sigma_j = 0$ for some $j \in \{1, \ldots, r\}$. By Theorem 1, we also have $\sigma_{j'}^2 = 0$ for all $j' \geq j$. This implies $\text{cost}(S_{j-1}) = 0$ by the second claim (using $k = j - 1$). This, in turn, implies that $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are contained in the $(j-1)$-dimensional subspace S_{j-1}, a contradiction. Hence $\sigma_j > 0$, proving the third claim. □

In light of Proposition 1, for each $j \in \{1, \ldots, r\}$, we define the unit vector

$$\mathbf{u}_j \;=\; \frac{1}{\sigma_j} \begin{bmatrix} \langle \mathbf{a}_1, \mathbf{v}_j \rangle \\ \vdots \\ \langle \mathbf{a}_m, \mathbf{v}_j \rangle \end{bmatrix} \;=\; \frac{1}{\sigma_j} A \mathbf{v}_j, \tag{2}$$

where A is the $m \times n$ matrix whose ith row is $\mathbf{a}_i^\mathsf{T}$. Note that $\sigma_j = \|A\mathbf{v}_j\|$ for each $j \in \{1, \ldots, r\}$. Proposition 1 also shows that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ form an ONB for the row space of A.

Proposition 2. *Consider the setting of Theorem 1 and Section 1.5. Then*

$$A \;=\; \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T}.$$

Proof. By the Unique Linear Transformation Theorem, it suffices to show that the left-hand side and right-hand side, as linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$, agree on a basis for $\mathbb{R}^n$.

We construct a basis as follows. We start with $\mathbf{v}_1, \ldots, \mathbf{v}_r$; these vectors form an ONB for the row space of A by Proposition 1. Let $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$ be an ONB for the nullspace of A. Since the row space and nullspace are orthogonal complements of each other, it follows that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form an ONB for $\mathbb{R}^n$.

Consider $i \in \{1, \ldots, r\}$. We have, by the definition of $\mathbf{u}_i$,

$$A\mathbf{v}_i \;=\; \sigma_i \mathbf{u}_i.$$

Moreover, since $\mathbf{v}_1, \ldots, \mathbf{v}_r$ are orthonormal, we have $\mathbf{v}_i^\mathsf{T} \mathbf{v}_j = 1$ if $i = j$ and 0 otherwise; so

$$\sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T} \mathbf{v}_i \;=\; \sigma_i \mathbf{u}_i$$

Now consider $i \in \{r+1, \ldots, n\}$. Since $\mathbf{v}_i$ is in the nullspace of A, we have $A\mathbf{v}_i = \mathbf{0}$. And since $\mathbf{v}_i$ is orthogonal to $\mathbf{v}_j$ for all $j \in \{1, \ldots, r\}$, we have $\sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T} \mathbf{v}_i = \mathbf{0}$. So we conclude that A and $\sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T}$ are the same matrices. □

There is one last property to establish, orthonormality of $\mathbf{u}_1, \ldots, \mathbf{u}_r$, as the next proposition shows.

Proposition 3. *Consider the setting of Theorem 1 and Section 1.5. Then* $\mathbf{u}_1, \ldots, \mathbf{u}_r$ *are orthonormal.*

The proof of Proposition 3 (see Appendix B) shows that if an ONB $\mathbf{v}_1, \ldots, \mathbf{v}_r$ for $\mathsf{CS}(A^\mathsf{T})$ doesn't lead to orthogonal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ via the definition in (2), then it is possible to improve the gain.

Since orthonormal vectors must be linearly independent, it readily follows from Proposition 2, Proposition 3, and the Basis Sufficiency Theorem that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ form an ONB for the column space of A.

Caution. It is possible that the choice of $\mathbf{v}_j$ in Line 3 of Algorithm 1 is not unique, i.e., there could be multiple unit vectors in $S_{j-1}^{\perp}$ that achieve the same gain. The analysis throughout Section 1 allows for any of the possibilities.

2 Singular value decomposition

2.1 Existence theorem

The analysis of Algorithm 1 in Sections 1.4 and 1.5 applies to any data set—and hence any $m \times n$ matrix A. So, the results there imply the following general theorem about arbitrary matrices.

Theorem 2 (Singular Value Decomposition Theorem)**.** *For any* $m \times n$ *matrix* A *with* $r = \mathrm{rank}(A)$*, there exist*

- *positive real numbers* $\sigma_1 \geq \cdots \geq \sigma_r$*,*
- *an ONB* $\mathbf{u}_1, \ldots, \mathbf{u}_r$ *for the column space of* A*, and*
- *an ONB* $\mathbf{v}_1, \ldots, \mathbf{v}_r$ *for the row space of* A*,*

such that

$$A \;=\; \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\intercal.$$

The decomposition of the matrix A shown in Theorem 2 is called a *singular value decomposition (SVD)* of A. The numbers $\sigma_1, \ldots, \sigma_r$ are called the *singular values* of A. The vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ are, respectively, the *left singular vector* and the *right singular vector* corresponding to the singular value σ_j. This decomposition expresses A as a sum of r outer products of vectors $\sigma_1 \mathbf{u}_1 \mathbf{v}_1^\intercal, \ldots, \sigma_r \mathbf{u}_r \mathbf{v}_r^\intercal$, each of which is an $m \times n$ matrix of rank 1.

The sequence of singular values is uniquely determined by the matrix A, but it is possible that there are multiple choices for corresponding singular vectors. For example, if A is the $n \times n$ identity matrix, then $\sigma_1 = \cdots = \sigma_n = 1$, but $\mathbf{v}_1, \ldots, \mathbf{v}_n$ can be any ONB for $\mathbb{R}^n$ (provided $\mathbf{u}_j = \mathbf{v}_j$ as well).

An SVD of $A^\intercal$ is obtained from an SVD of A, except switching the roles of left- and right-singular vectors.

2.2 Geometric interpretation: two-dimensional case

Using an SVD, we can interpret the behavior of A, as a linear transformation $T_A \colon \mathbb{R}^n \to \mathbb{R}^m$ given by $T_A(\mathbf{x}) = A\mathbf{x}$, when applied to the unit sphere (i.e., the set of all unit vectors) in $\mathbb{R}^n$. We first start with the case where $m = n = 2$ and $\mathrm{rank}(A) = 2$. The unit sphere in $\mathbb{R}^2$ is simply the unit circle.

Adopt the notations from Theorem 2 for an SVD of A, and let $\mathcal{U} = (\mathbf{u}_1, \mathbf{u}_2)$ and $\mathcal{V} = (\mathbf{v}_1, \mathbf{v}_2)$ be, respectively, the ONB's corresponding to the left and right singular vectors. Observe that

$$T_A(\mathbf{v}_j) \;=\; \sigma_j \mathbf{u}_j \quad \text{for each } j \in \{1, 2\}. \tag{3}$$

A vector $\mathbf{x} \in \mathbb{R}^2$ has coordinates $[\mathbf{x}]_{\mathcal{V}} = ([\mathbf{x}]_{\mathcal{V},1}, [\mathbf{x}]_{\mathcal{V},2})$ with respect to $\mathcal{V}$:

$$\mathbf{x} \;=\; [\mathbf{x}]_{\mathcal{V},1}\, \mathbf{v}_1 + [\mathbf{x}]_{\mathcal{V},1}\, \mathbf{v}_2.$$

Let $\mathbf{y} = T_A(\mathbf{x})$. By linearity of the transformation and (3),

$$\begin{aligned} \mathbf{y} \;=\; T_A(\mathbf{x}) \;&=\; T_A([\mathbf{x}]_{\mathcal{V},1}\, \mathbf{v}_1 + [\mathbf{x}]_{\mathcal{V},1}\, \mathbf{v}_2) \\ &=\; [\mathbf{x}]_{\mathcal{V},1}\, T_A(\mathbf{v}_1) + [\mathbf{x}]_{\mathcal{V},1}\, T_A(\mathbf{v}_2) \\ &=\; [\mathbf{x}]_{\mathcal{V},1}\, \sigma_1 \mathbf{u}_1 + [\mathbf{x}]_{\mathcal{V},1}\, \sigma_2 \mathbf{u}_2. \end{aligned}$$

So the coordinates of $\mathbf{y}$ with respect to $\mathcal{U}$ are

$$[\mathbf{y}]_{\mathcal{U}} \;=\; \begin{bmatrix} [\mathbf{y}]_{\mathcal{U},1} \\ [\mathbf{y}]_{\mathcal{U},2} \end{bmatrix} \;=\; \begin{bmatrix} [\mathbf{x}]_{\mathcal{V},1}\,\sigma_1 \\ [\mathbf{x}]_{\mathcal{V},2}\,\sigma_2 \end{bmatrix} \;=\; \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} [\mathbf{x}]_{\mathcal{V},1} \\ [\mathbf{x}]_{\mathcal{V},2} \end{bmatrix}.$$

In other words, the matrix representation $[T_A]_{\mathcal{V}\to\mathcal{U}}$ of T_A with respect to the input space basis $\mathcal{V}$ and output space basis $\mathcal{U}$ is a diagonal matrix. Observe that we have the relation

$$[\mathbf{x}]_{\mathcal{V},j} \;=\; \frac{[\mathbf{y}]_{\mathcal{U},j}}{\sigma_j} \quad \text{for each } j \in \{1,2\}.$$

If $\mathbf{x}$ is a unit vector, then $[\mathbf{x}]_{\mathcal{V},1}^2 + [\mathbf{x}]_{\mathcal{V},2}^2 = \|\mathbf{x}\|^2 = 1$ by Parseval's identity, and therefore $\mathbf{y} = T_A(\mathbf{x})$ must satisfy

$$\frac{[\mathbf{y}]_{\mathcal{U},1}^2}{\sigma_1^2} + \frac{[\mathbf{y}]_{\mathcal{U},2}^2}{\sigma_2^2} \;=\; 1.$$

This is the equation for an ellipse with major axis $\mathbf{u}_1$ and minor axis $\mathbf{u}_2$. (It is again a circle if $\sigma_1 = \sigma_2$, in which case there are no distinguished axes.)

2.3 Geometric interpretation: general case

We now consider the general case of what we set out to do in Section 2.2. We may regard A as a linear transformation between the r-dimensional subspaces $\mathsf{CS}(A^\intercal)$ and $\mathsf{CS}(A)$, again given by $T_A(\mathbf{x}) = A\mathbf{x}$. This is because any part of an n-vector $\mathbf{x}$ in the nullspace of A is "nullified" (i.e., mapped to zero) by T_A, and all that matters is the part of $\mathbf{x}$ in $\mathsf{CS}(A^\intercal)$.

Let $\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ and $\mathcal{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, respectively, be ordered ONB's for $\mathsf{CS}(A^\intercal)$ and $\mathsf{CS}(A)$, formed by left and right singular vectors of A. Then, following the same line of reasoning as in Section 2.2, we find that $[T_A]_{\mathcal{V}\to\mathcal{U}}$ is a diagonal matrix:

$$[T_A]_{\mathcal{V}\to\mathcal{U}} \;=\; \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}.$$

(The fact that the matrix representation is diagonal is what makes the bases $\mathcal{V}$ and $\mathcal{U}$ "nice" to use when working with T_A.[2]) If $\mathbf{x}$ is a unit vector in $\mathsf{CS}(A^\intercal)$, then $\mathbf{y} = T_A(\mathbf{x})$ must satisfy

$$\frac{[\mathbf{y}]_{\mathcal{U},1}^2}{\sigma_1^2} + \cdots + \frac{[\mathbf{y}]_{\mathcal{U},r}^2}{\sigma_r^2} \;=\; 1,$$

which is the equation for an ellipsoid in $\mathsf{CS}(A)$ with axes specified by $\mathbf{u}_1, \ldots, \mathbf{u}_r$. If $r < m$, then this ellipsoid is a "degenerate" ellipsoid in $\mathbb{R}^m$.

2.4 Compact SVD

An SVD of a matrix A is sometimes expressed as a factorization of A into a product of three matrices:

$$A \;=\; U \Sigma V^\intercal,$$

where

$$U \;=\; \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r \\ \downarrow & & \downarrow \end{bmatrix}}_{m \times r \text{ matrix}}, \quad \Sigma \;=\; \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}}_{r \times r \text{ diagonal matrix}}, \quad V \;=\; \underbrace{\begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{v}_1 & \cdots & \mathbf{v}_r \\ \downarrow & & \downarrow \end{bmatrix}}_{n \times r \text{ matrix}}.$$

This factorization is called a *compact SVD* of A.[3] The equivalence of this factorization of A and the additive decomposition of A from Theorem 2 can be checked by using the "sum of outer products" view of matrix multiplication. The diagonal matrix Σ is the matrix representation of T_A from Section 2.3 with respect to the ONB's given by the left and right singular vectors.

The matrices $U^\intercal$ and V transform the matrix A into Σ via matrix multiplication on the left and right. To see this, observe that $U^\intercal U = I$ because the left singular vectors are orthonormal, and $V^\intercal V = I$ because the right singular vectors are orthonormal. Therefore

$$U^\intercal A V \;=\; U^\intercal U \Sigma V^\intercal V \;=\; \Sigma.$$

[2] If A is an $n \times n$ (square) matrix, then one could hope to find a single ordered basis $\mathcal{V}$ for $\mathbb{R}^n$ such that $[T_A]_{\mathcal{V} \to \mathcal{V}}$ is diagonal. This is not always possible, but it is in some important cases, as we'll see later.

[3] Compare this to a "full SVD", described in Appendix C.

2.5 Truncated SVD

A *rank-k truncated SVD* of a matrix A obtained by retaining the part of an SVD of A corresponding to k largest singular values. In the notation of Theorem 2, it leads to a *rank-k SVD approximation* of A, defined by:

$$\hat{A} \;=\; \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^\mathsf{T}.$$

If $k \geq \mathrm{rank}(A)$, then this rank-k approximation is exactly the same as A, i.e., $\hat{A} = A$. But if $k < \mathrm{rank}(A)$, it is only an approximation of A.

Note that $\hat{A}$ is an $m \times n$ matrix. However, it is represented by just $(m+n+1)k$ numbers, which can be much smaller than mn when k is small. Computing the matrix-vector multiplication $\hat{A}\mathbf{x}$ for a given n-vector $\mathbf{x}$ also requires only $(m+n+1)k$ scalar multiplications (as opposed to mn for $A\mathbf{x}$, in general). The computational savings are significant when m and n are large.

The approximation quality, suitably defined, achieved by a rank-k truncated SVD is optimal: there is no better rank-k matrix that achieves the same approximation quality. This is the content of the following theorem, which is closely related to the optimality of Algorithm 1 for k-BFS in Theorem 1.[4]

Theorem 3 (Eckart-Young Theorem)**.** *Let A be an $m \times n$ matrix A of rank r, and fix any $k \in \{0, \ldots, r\}$. Let $\hat{A}$ be a rank-k SVD approximation of A. Then, for any other $m \times n$ matrix B of rank k,*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{i,j} - B_{i,j})^2 \;\geq\; \sum_{i=1}^{m}\sum_{j=1}^{n}(A_{i,j} - \hat{A}_{i,j})^2 \;=\; \sigma_{k+1}^2 + \cdots + \sigma_r^2,$$

where $\sigma_1 \geq \cdots \geq \sigma_r$ are the singular values of A.

Example. Consider the 4×4 matrix

$$A \;=\; \begin{bmatrix} 3 & 1 & 2 & 0 \\ -1 & -3 & 0 & -2 \\ 0 & 2 & 1 & 3 \\ -2 & 0 & -3 & -1 \end{bmatrix}.$$

[4] This theorem has a complex history.

The first two components of an SVD are

$$[\mathbf{u}_1, \mathbf{u}_2] \;=\; \begin{bmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \\ 1/2 & -1/2 \\ -1/2 & -1/2 \end{bmatrix}, \quad \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \;=\; \begin{bmatrix} 6 & \\ & 4 \end{bmatrix}, \quad [\mathbf{v}_1, \mathbf{v}_2] \;=\; \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}.$$

The resulting rank-2 approximation is

$$\hat{A} \;=\; \begin{bmatrix} 5/2 & 1/2 & 5/2 & 1/2 \\ -1/2 & -5/2 & -1/2 & -5/2 \\ 1/2 & 5/2 & 1/2 & 5/2 \\ -5/2 & -1/2 & -5/2 & -1/2 \end{bmatrix}.$$

Every entry in $\hat{A}$ differs from the corresponding entry of A by $\pm 1/2$. The sum of the squared differences is $\sum_{i=1}^{4} \sum_{j=1}^{4} (A_{i,j} - \hat{A}_{i,j})^2 = 4$. ∎

2.6 Pseudoinverse

A compact SVD $U\Sigma V^\intercal$ of an $m \times n$ matrix A defines an "inverse"-like matrix called the *(Moore-Penrose) pseudoinverse* of A:

$$A^+ \;=\; V\Sigma^{-1}U^\intercal.$$

Notice that $\mathsf{CS}(A^+) = \mathsf{CS}(A^\intercal)$, and $\mathsf{CS}((A^+)^\intercal) = \mathsf{CS}(A)$.[5] The Moore-Penrose pseudoinverse behaves almost like an inverse of A in the following sense:

$$\begin{aligned} AA^+ \;&=\; U\Sigma V^\intercal V\Sigma^{-1}U^\intercal \;=\; UU^\intercal, \\ A^+A \;&=\; V\Sigma^{-1}U^\intercal U\Sigma V^\intercal \;=\; VV^\intercal, \end{aligned}$$

which are, respectively, the orthoprojectors for the column space and row space of A. In general, these orthoprojectors need not be an identity matrix. But:

- If $\mathrm{rank}(A) = m$, then $\mathsf{CS}(A) = \mathbb{R}^m$, and $AA^+ = UU^\intercal = I$.
- If $\mathrm{rank}(A) = n$, then $\mathsf{CS}(A^\intercal) = \mathbb{R}^n$, and $A^+A = VV^\intercal = I$.

[5] Sometimes we write $A^{+\intercal}$ to mean $(A^+)^\intercal$.

3 Application to least squares approximation

One important use of the SVD, and in particular, the Moore-Penrose pseudoinverse, is for solving the least squares approximation problem. Recall that, in that problem, one is given an $n \times p$ matrix A and an n-vector $\mathbf{b}$, and the goal is to find a p-vector $\mathbf{x}$ that minimizes $\|A\mathbf{x} - \mathbf{b}\|^2$. We saw that any solution to the system of linear equations called the normal equations,

$$(A^\intercal A)\mathbf{x} \;=\; A^\intercal \mathbf{b},$$

yields a minimizer of $\|A\mathbf{x}-\mathbf{b}\|^2$. But it is possible that the normal equations do not have a unique solution; this is the case if $\mathrm{rank}(A) < p$.

A particular solution is given by the Moore-Penrose pseudoinverse: $A^+\mathbf{b}$. To verify that $A^+\mathbf{b}$ is a solution to the normal equations, recall that solutions to the normal equations are the same as solutions to $A\mathbf{x} = \mathbf{b}_0$, where $\mathbf{b}_0$ is the orthogonal projection of $\mathbf{b}$ to $\mathsf{CS}(A)$. But AA^+ is the orthoprojector for $\mathsf{CS}(A)$, so $A(A^+\mathbf{b}) = (AA^+)\mathbf{b} = \mathbf{b}_0$.

But what is special about $A^+\mathbf{b}$?

- It is the unique solution in the row space of A. To see this, recall that since $\mathbf{b}_0 \in \mathsf{CS}(A)$, there is a unique solution to $A\mathbf{x} = \mathbf{b}_0$ contained in $\mathsf{CS}(A^\intercal)$. Since $\mathsf{CS}(A^+) = \mathsf{CS}(A^\intercal)$, and $A^+\mathbf{b}$ is a solution to $A\mathbf{x} = \mathbf{b}_0$, it follows that $A^+\mathbf{b}$ is the only solution to the normal equations in $\mathsf{CS}(A^\intercal)$.

- It is the unique solution of smallest Euclidean norm. To see this, recall that any solution $\mathbf{x}$ to $A\mathbf{x} = \mathbf{b}_0$ can be written as the sum of a particular solution—say, $A^+\mathbf{b}$—and a vector $\mathbf{z}$ in the nullspace $\mathsf{NS}(A)$. If $\mathbf{z} \neq \mathbf{0}$, then $\|\mathbf{z}\| > 0$ by positive definiteness of the norm. Since $\mathsf{CS}(A^\intercal)$ and $\mathsf{NS}(A)$ are orthogonal complements, the Pythagorean Theorem implies

  $$\|\mathbf{x}\|^2 \;=\; \|A^+\mathbf{b}\|^2 + \|\mathbf{z}\|^2,$$

  which is strictly larger than $\|A^+\mathbf{b}\|^2$ since $\|\mathbf{z}\|^2 > 0$.

4 Positive semidefinite matrices

A very important class of matrices is the class of positive semidefinite matrices.[6] We say an $n \times n$ matrix A is *positive semidefinite (PSD)* if there is another matrix B such that $A = B^\intercal B$. In fact, an SVD of A can be obtained from an SVD of B.

Proposition 4. *Suppose $A = B^\intercal B$, and $B = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\intercal$ is a singular value decomposition of B. Then $A = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^\intercal$ is a singular value decomposition of A.*

Proof. Let $B = U\Sigma V^\intercal$ be the compact SVD of B, formed from the SVD given in the hypothesis in the usual way. Then

$$\begin{aligned} A &= B^\intercal B \\ &= (V\Sigma U^\intercal)(U\Sigma V^\intercal) \\ &= V\Sigma\Sigma V^\intercal && (\text{since } U^\intercal U = I) \\ &= V\Sigma^2 V^\intercal && (\text{since } \Sigma \text{ is diagonal}). \end{aligned}$$

So $A = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^\intercal$, which is a singular value decomposition of A where $\sigma_1^2, \ldots, \sigma_r^2$ are the singular values, and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ serve as both the corresponding left singular vectors and the corresponding right singular vectors. □

PSD matrices arise in many contexts. We have already seen them come up in the normal equations for least squares. They are also important in multivariate statistics: covariance matrices are always of the form $B^\intercal B$ (where B is a "data matrix" whose rows have been "centered").

There is an equivalent characterization of positive semidefiniteness: an $n \times n$ matrix A is positive semidefinite if it is symmetric and $\langle \mathbf{x}, A\mathbf{x} \rangle$ for all n-vectors $\mathbf{x}$. (This is, in fact, the more common definition of a positive semidefinite matrix.)

Theorem 4. *The following statements about an $n \times n$ matrix A are equivalent.*

1. *There is a matrix B such that $A = B^\intercal B$.*

[6] Just like many terms in mathematics (e.g., "orthogonal"), the term "positive semidefinite" is applied in many different contexts. We have previous defined a very closely related property, "positive definiteness", for inner products and norms; later, we will also define "positive definite matrices" as a special case of "positive semidefinite matrices".

2. A is symmetric and $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all n-vectors $\mathbf{x}$.

Proof. First, suppose there is a matrix B such that $A = B^\intercal B$. We will show that A is symmetric and $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all n-vectors $\mathbf{x}$. For any $\mathbf{x}$,

$$\begin{aligned} \langle \mathbf{x}, A\mathbf{x} \rangle &= \langle \mathbf{x}, B^\intercal B\mathbf{x} \rangle && \text{(plugging in } A = B^\intercal B) \\ &= \langle B\mathbf{x}, B\mathbf{x} \rangle && \text{(using } \langle \mathbf{u}, M\mathbf{v} \rangle = \langle M^\intercal \mathbf{u}, \mathbf{v} \rangle) \\ &\geq 0 && \text{(by positive definiteness of inner product).} \end{aligned}$$

Now, let us instead suppose A is symmetric and $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all n-vectors $\mathbf{x}$. We will show that there is a matrix B such that $A = B^\intercal B$. Consider an SVD $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i$ of A. Observe that if $\mathbf{u}_i = \mathbf{v}_i$ for all $i \in \{1, \ldots, r\}$, then we would be able to construct the desired matrix B:

$$B = \begin{bmatrix} \longleftarrow & \sqrt{\sigma_1}\mathbf{v}_1^\intercal & \longrightarrow \\ & \vdots & \\ \longleftarrow & \sqrt{\sigma_r}\mathbf{v}_r^\intercal & \longrightarrow \end{bmatrix}.$$

In terms of the corresponding compact SVD $A = U\Sigma V^\intercal$ of A, this matrix B can be written as $B = \sqrt{\Sigma} V^\intercal$, where $\sqrt{\Sigma}$ is the diagonal matrix whose diagonal entries are the square roots of that of Σ. Indeed, if $\mathbf{u}_i = \mathbf{v}_i$ for all $i \in \{1, \ldots, r\}$, then $U = V$, and

$$B^\intercal B = V\sqrt{\Sigma}\sqrt{\Sigma}V^\intercal = V\Sigma V^\intercal = A.$$

So, we just need to prove that $\mathbf{u}_i = \mathbf{v}_i$ for all $i \in \{1, \ldots, r\}$: this is the content of Proposition 5. □

Proposition 5. *Suppose A is an $n \times n$ symmetric matrix such that $\langle \mathbf{x}, A\mathbf{x} \rangle \geq 0$ for all n-vectors $\mathbf{x}$, and $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\intercal$ is an SVD of A. Then $\mathbf{u}_i = \mathbf{v}_i$ for all $i \in \{1, \ldots, r\}$.*

Proof. Observe that $A^2 = AA = A^\intercal A$ because A is symmetric. Therefore, by Proposition 4, $A^2 = \sum_{i=1}^r \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^\intercal$ is an SVD of A^2. Now consider any $i \in \{1, \ldots, r\}$, and define $\mathbf{x} = A\mathbf{v}_i - \sigma_i \mathbf{v}_i$. From the SVD of A, we see that $A\mathbf{v}_i = \sigma_i \mathbf{u}_i$, so $\mathbf{x} = \sigma_i(\mathbf{u}_i - \mathbf{v}_i)$. Since $\sigma_i \neq 0$, it follows that $\mathbf{x} = \mathbf{0}$ if and only

if $\mathbf{u}_i = \mathbf{v}_i$. Consider the following sequence of inequalities and equalities:

$$\begin{aligned}
0 &\leq \langle \mathbf{x}, A\mathbf{x} \rangle && \text{(by hypothesis)} \\
&= \langle \mathbf{x}, A(A\mathbf{v}_i - \sigma_i \mathbf{v}_i) \rangle \\
&= \langle \mathbf{x}, A^2\mathbf{v}_i - \sigma_i A\mathbf{v}_i \rangle && \text{(linearity)} \\
&= \langle \mathbf{x}, \sigma_i^2 \mathbf{v}_i - \sigma_i A\mathbf{v}_i \rangle && \text{(using SVD of } A^2\text{)} \\
&= -\sigma_i \langle \mathbf{x}, -\sigma_i \mathbf{v}_i + A\mathbf{v}_i \rangle && \text{(linearity)} \\
&= -\sigma_i \langle \mathbf{x}, \mathbf{x} \rangle \\
&\leq 0
\end{aligned}$$

where the last step follows by the positive definiteness of inner product. This shows that $\sigma_i \langle \mathbf{x}, \mathbf{x} \rangle = 0$. Again, by positive definiteness of inner product, we conclude that $\mathbf{x} = \mathbf{0}$, and therefore $\mathbf{u}_i = \mathbf{v}_i$. □

A special class of positive semidefinite matrices are the positive definite matrices. An $n \times n$ matrix A is *positive definite* if A is positive semidefinite, and $\langle \mathbf{x}, A\mathbf{x} \rangle = 0$ implies that $\mathbf{x} = \mathbf{0}$.

Proposition 6. *A square matrix is positive definite if and only if it is positive semidefinite and invertible.*

Proof. Suppose a square matrix A is not invertible. By the Invertibility Theorem, the nullspace of A has a non-zero vector $\mathbf{x}$. Then $\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{0} \rangle = 0$, and yet $\mathbf{x} \neq \mathbf{0}$. So A is not positive definite.

Now suppose the $n \times n$ matrix A is positive semidefinite and invertible. Let $A = U\Sigma V^\intercal$ be a compact SVD of A. By the Invertibility Theorem, the rank of A is n, and by Proposition 5, we have $U = V$. Therefore V is orthogonal, and $\langle \mathbf{x}, A\mathbf{x} \rangle = \langle \mathbf{x}, V\Sigma V^\intercal \mathbf{x} \rangle = \langle \sqrt{\Sigma} V^\intercal \mathbf{x}, \sqrt{\Sigma} V^\intercal \mathbf{x} \rangle$, which is zero only if $\sqrt{\Sigma} V^\intercal \mathbf{x} = \mathbf{0}$ by positive definiteness of inner product. (Here, $\sqrt{\Sigma}$ is the diagonal matrix whose diagonal entries are the square roots of that of Σ.) But observe that both $\sqrt{\Sigma}$ and $V^\intercal$ are invertible, so their product is also invertible. By the Invertibility Theorem, $\sqrt{\Sigma} V^\intercal \mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. Hence A is positive definite. □

A Finding the best fitting line

The following algorithm approximately solves BFL problem.

Algorithm 2 Power method for best fitting line

Input: $m \times n$ matrix A; initial vector $\mathbf{x}_0 \in \mathsf{CS}(A^\intercal)$; number of iterations t.

1: **for** $s = 1, \ldots, t$ **do**
2: Let $\mathbf{y}_s = A\mathbf{x}_{s-1}$.
3: Let $\mathbf{x}_s = A^\intercal \mathbf{y}_s$.
4: **end for**
5: **return** Unit vector $\mathbf{z} = \mathbf{x}_t / \|\mathbf{x}_t\|$.

Theorem 5. *Let A be an $m \times n$ matrix. Let σ_1 be the largest singular value of A, and let $\mathbf{v}_1$ be a left singular vector corresponding to σ_1. Let $\mathbf{x}_0 \in \mathsf{CS}(A^\intercal)$ be a non-zero vector with $\langle \mathbf{v}_1, \mathbf{x}_0 \rangle^2 \geq \delta \|\mathbf{x}_0\|^2$ for some $\delta \in (0, 1]$. For any $\varepsilon \in (0, 1)$, if the number of iterations t satisfies*

$$t \;\geq\; \frac{1}{2\varepsilon} \ln \frac{1}{\varepsilon\delta},$$

then the execution of Algorithm 2 on A and initial vector $\mathbf{x}_0$ for t iterations returns a unit vector $\mathbf{z}$ satisfying

$$\|A\mathbf{z}\|^2 \;\geq\; \frac{1-\varepsilon}{1+\varepsilon}\, \sigma_1^2.$$

Note that the guarantee for Algorithm 2 in Theorem 5 requires an initial non-zero vector $\mathbf{x}_0 \in \mathsf{CS}(A^\intercal)$ that satisfies $\langle \mathbf{v}_1, \mathbf{x}_0 \rangle^2 \geq \delta \|\mathbf{x}_0\|^2$ for some $\delta \in (0, 1]$. To obtain a vector in $\mathsf{CS}(A^\intercal)$, it suffices to multiply $A^\intercal$ by any m-vector. Or, just choose a vector $\mathbf{x}' \in \mathbb{R}^n$ and regard $\mathbf{x}_0$ as the orthogonal projection of $\mathbf{x}'$ to $\mathsf{CS}(A^\intercal)$; the first iteration nullifies the part of $\mathbf{x}'$ in $\mathsf{NS}(A)$.

However, not just any vector in $\mathsf{CS}(A^\intercal)$ is good enough. We want to use Algorithm 2 with a non-zero $\mathbf{x}_0$ such that $\langle \mathbf{v}_1, \mathbf{x}_0 \rangle^2 / \|\mathbf{x}_0\|^2$ is not too small. One way to do this is to choose a unit vector (in $\mathbb{R}^n$) uniformly at random. It is unlikely for $\langle \mathbf{v}_1, \mathbf{x}_0 \rangle^2 / \|\mathbf{x}_0\|^2$ to be much smaller than $1/n$. If the unlikely event does not happen, then the requirement on t is on the order of $\log n$ (regarding ε as a constant).[7]

[7] If you're not feeling lucky, just run Algorithm 2 a half dozen times or so with different (randomly chosen) initial vectors, and take the best result. This dramatically decreases the chance of failure.

Proof of Theorem 5. Notice that $\|A\mathbf{z}\|^2 = \|A\mathbf{x}_t\|^2/\|\mathbf{x}_t\|^2 = \|\mathbf{y}_{t+1}\|^2/\|\mathbf{x}_t\|^2$. So, to obtain a lower-bound on this quantity, we will obtain a lower-bound on $\|\mathbf{y}_{t+1}\|^2$ and an upper-bound on $\|\mathbf{x}_t\|^2$.

Denote the singular values of A by $\sigma_1 \geq \cdots \geq \sigma_r > 0$, where r is the rank of A. Let $\mathcal{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ and $\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ be ordered ONB's composed of, respectively, left singular vectors and right singular vectors corresponding to the singular values. Let ℓ denote the number of singular values σ_j satisfying $\sigma_j^2 \geq (1-\varepsilon)\sigma_1^2$. The assumption on $\mathbf{x}_0$ (via Parseval's identity) ensures

$$[\mathbf{x}_0]_{\mathcal{V},1}^2 \;\geq\; \delta\|\mathbf{x}_0\|^2 \;=\; \delta\sum_{j=1}^{r}[\mathbf{x}_0]_{\mathcal{V},j}^2 \;\geq\; \delta\sum_{j=\ell+1}^{r}[\mathbf{x}_0]_{\mathcal{V},j}^2,$$

where $[\mathbf{v}]_{\mathcal{V}} = ([\mathbf{v}]_{\mathcal{V},1}, \ldots, [\mathbf{v}]_{\mathcal{V},r})$. The assumption on t (via the inequality $1 + a \leq e^a$ for any real number a) ensures that

$$(1-\varepsilon)^{2t} \;\leq\; e^{-2\varepsilon t} \;\leq\; \varepsilon\delta.$$

Using coordinates with respect to the ordered bases $\mathcal{U}$ and $\mathcal{V}$, the updates in iteration s of Algorithm 2 are

$$[\mathbf{y}_s]_{\mathcal{U}} \;=\; \Sigma[\mathbf{x}_{s-1}]_{\mathcal{V}} \qquad \text{and} \qquad [\mathbf{x}_s]_{\mathcal{V}} \;=\; \Sigma[\mathbf{y}_s]_{\mathcal{U}} \;=\; \Sigma^2[\mathbf{x}_{s-1}]_{\mathcal{V}},$$

where Σ is the $r \times r$ diagonal matrix of singular values $\sigma_1, \ldots, \sigma_r$. So, by a simple induction, we have

$$[\mathbf{x}_t]_{\mathcal{V}} \;=\; \Sigma^{2t}[\mathbf{x}_0]_{\mathcal{V}} \qquad \text{and} \qquad [\mathbf{y}_{t+1}]_{\mathcal{U}} \;=\; \Sigma^{2t+1}[\mathbf{x}_0]_{\mathcal{V}}.$$

Using Parseval's identity and the choice of ℓ, we obtain

$$\begin{aligned}
\|\mathbf{y}_{t+1}\|^2 \;=\; \|[\mathbf{y}_{t+1}]_{\mathcal{U}}\|^2 \;&=\; \|\Sigma^{2t+1}[\mathbf{x}_0]_{\mathcal{V}}\|^2 \\
&=\; \sum_{j=1}^{r}\sigma_j^{4t+2}[\mathbf{x}_0]_{\mathcal{V},j}^2 \\
&\geq\; \sum_{j=1}^{\ell}\sigma_j^{4t+2}[\mathbf{x}_0]_{\mathcal{V},j}^2 \\
&\geq\; (1-\varepsilon)\sigma_1^2\sum_{j=1}^{\ell}\sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2,
\end{aligned}$$

and also, using the conditions on t and δ,

$$\begin{aligned}
\|\mathbf{x}_t\|^2 \;=\; \|[\mathbf{x}_t]_{\mathcal{V}}\|^2 \;&=\; \|\Sigma^{2t}[\mathbf{x}_0]_{\mathcal{V}}\|^2 \\
&=\; \sum_{j=1}^{r} \sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2 \\
&\leq\; \sum_{j=1}^{\ell} \sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2 + (1-\varepsilon)^{2t}\sigma_1^{4t}\sum_{j=\ell+1}^{r}[\mathbf{x}_0]_{\mathcal{V},j}^2 \\
&\leq\; \sum_{j=1}^{\ell} \sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2 + \varepsilon\delta\sigma_1^{4t}\sum_{j=\ell+1}^{r}[\mathbf{x}_0]_{\mathcal{V},j}^2 \quad \text{(by condition on } t\text{)} \\
&\leq\; \sum_{j=1}^{\ell} \sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2 + \varepsilon\sigma_1^{4t}[\mathbf{x}_0]_{\mathcal{V},1}^2 \quad \text{(by condition on } \delta\text{)} \\
&\leq\; (1+\varepsilon)\sum_{j=1}^{\ell} \sigma_j^{4t}[\mathbf{x}_0]_{\mathcal{V},j}^2.
\end{aligned}$$

So, the ratio $\|\mathbf{y}_{t+1}\|^2/\|\mathbf{x}_t\|^2$ satisfies

$$\frac{\|\mathbf{y}_{t+1}\|^2}{\|\mathbf{x}_t\|^2} \;\geq\; \frac{1-\varepsilon}{1+\varepsilon}\,\sigma_1^2. \qquad \square$$

B Orthogonality of the left singular vectors

Proof of Proposition 3. Suppose for sake of contradiction that there is some pair (i,j) with $1 \leq i < j \leq r$ such that $\langle \mathbf{u}_i, \mathbf{u}_j \rangle \neq 0$. Take the pair (i,j) for which i is as small as possible. Define $c = \langle \mathbf{u}_i, \mathbf{u}_j \rangle$ and $\delta = \sigma_j/\sigma_i$. By assumption, $c \neq 0$, and by Proposition 1, $\delta > 0$. We define the unit vector

$$\mathbf{w}_i \;=\; \frac{1}{\|\mathbf{v}_i + \delta c \mathbf{v}_j\|}\,(\mathbf{v}_i + \delta c \mathbf{v}_j).$$

It suffices to prove that $\mathbf{w}_i$ has higher gain than $\mathbf{v}_i$ among unit vectors in $S_{i-1}^{\perp}$, a contradiction, which then leads us to conclude that no such pair (i,j) exists, i.e., that $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal.

The main quantity to study is $\mathbf{u}_i^\top A\mathbf{w}_i$; we will obtain bounds on this quantity from above and below. We start with the lower-bound. Observe that by linearity and the definitions of $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$, c, and δ,

$$\begin{aligned}
\mathbf{u}_i^\top A\mathbf{w}_i &= \mathbf{u}_i^\top A\left(\frac{1}{\|\mathbf{v}_i + \delta c\mathbf{v}_j\|}\left(\mathbf{v}_i + \delta c\mathbf{v}_j\right)\right) \\
&= \mathbf{u}_i^\top\left(\frac{1}{\|\mathbf{v}_i + \delta c\mathbf{v}_j\|}\left(\sigma_i\mathbf{u}_i + \delta c\sigma_j\mathbf{u}_j\right)\right) \\
&= \frac{\sigma_i + \delta c^2\sigma_j}{\|\mathbf{v}_i + \delta c\mathbf{v}_j\|} \\
&= \frac{1 + \delta^2 c^2}{\|\mathbf{v}_i + \delta c\mathbf{v}_j\|}\,\sigma_i.
\end{aligned}$$

By the orthonormality of $\mathbf{v}_1, \ldots, \mathbf{v}_r$ and the Pythagorean Theorem,

$$\|\mathbf{v}_i + \delta c\mathbf{v}_j\| \;=\; \sqrt{\|\mathbf{v}_i\|^2 + \|\delta c\mathbf{v}_j\|^2} \;=\; \sqrt{1 + \delta^2 c^2}.$$

Therefore, using $\delta^2 c^2 > 0$,

$$\mathbf{u}_i^\top A\mathbf{w}_i \;=\; \frac{1 + \delta^2 c^2}{\sqrt{1 + \delta^2 c^2}}\,\sigma_i \;=\; \sqrt{1 + \delta^2 c^2}\,\sigma_i \;>\; \sigma_i \;=\; \sqrt{\mathrm{gain}(\mathbf{v}_i)}.$$

Now we obtain a upper-bound on $\mathbf{u}_i^\top A\mathbf{w}_i$. Using the Cauchy-Schwarz Inequality and the fact $\|\mathbf{u}_i\| = 1$, we have

$$\mathbf{u}_i^\top A\mathbf{w}_i \;\leq\; \|\mathbf{u}_i\|\|A\mathbf{w}_i\| \;=\; \|A\mathbf{w}_i\| \;=\; \sqrt{\mathrm{gain}(\mathbf{w}_i)}.$$

Combining the upper- and lower-bounds on $\mathbf{u}_i^\top A\mathbf{w}_i$, we arrive at

$$\mathrm{gain}(\mathbf{w}_i) \;>\; \mathrm{gain}(\mathbf{v}_i).$$

But $\mathbf{w}_i$ is a linear combination of $\mathbf{v}_i$ and $\mathbf{v}_j$, both of which are in $S_{i-1}^\perp$ by construction and the fact $i < j$; hence $\mathbf{w}_i \in S_{i-1}^\perp$ as well. So the inequality above contradicts the fact that $\mathbf{v}_i$ has the highest gain among unit vectors in $S_{i-1}^\perp$ (as per Line 3 in Algorithm 1). This completes the proof. □

C Full SVD

A *full SVD* is a factorization of A into the product of three matrices, $A = U\Sigma V^\intercal$, but the matrices are defined somewhat differently compared to how they are defined in a compact SVD. To obtain a full SVD of A:

- Again, start with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of A, along with corresponding left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and right singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$.
- Extend the ONB $\mathbf{u}_1, \ldots, \mathbf{u}_r$ to an ONB for $\mathbb{R}^m$ to get $\mathbf{u}_1, \ldots, \mathbf{u}_m$. This is achieved by augmenting $\mathbf{u}_1, \ldots, \mathbf{u}_r$ with a basis for the left nullspace of A.
- Similarly, we extend the ONB $\mathbf{v}_1, \ldots, \mathbf{v}_r$ to an ONB for $\mathbb{R}^n$, to get $\mathbf{v}_1, \ldots, \mathbf{v}_n$.
- Then $U = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ is an $m \times m$ orthogonal matrix, and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ is an $n \times n$ orthogonal matrix.
- To write $A = U\Sigma V^\intercal$, the middle matrix Σ must be $m \times n$, and this is obtained just by extending the $r \times r$ diagonal matrix of singular values to a $m \times n$ matrix, filling in with zeros as needed:[8]

$$\Sigma = \underbrace{\left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & \text{zeros} \\ & & \sigma_r & \\ \hline & \text{zeros} & & \text{zeros} \end{array}\right]}_{m \times n \text{ matrix}}.$$

  There are $\min\{m, n\}$ "diagonal" entries in Σ; the first r of them are $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and the remaining $\min\{m, n\} - r$ of them are $\sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0$. The rank of A is simply the number of non-zero σ_i's.

[8] It is possible that some of these blocks of zeros are empty. For example, if $r = m < n$, then the bottom two blocks of zeros would not appear.

D Eckart-Young Theorem

Proof of Theorem 3. Let $\sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\mathsf{T}$ be an SVD of A, and suppose the rank-k approximation of A is obtained be retaining the part of this decomposition corresponding to the k largest singular values $\sigma_1, \ldots, \sigma_k$. Let $S_k = \mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})$ and let $P_k = \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\mathsf{T}$ be the orthoprojector for S_k. Then observe that

$$AP_k \;=\; \left(\sum_{i=1}^r \sigma_i \, \mathbf{u}_i \mathbf{v}_i^\mathsf{T}\right)\left(\sum_{j=1}^k \mathbf{v}_j \mathbf{v}_j^\mathsf{T}\right) \;=\; \sum_{i=1}^r \sum_{j=1}^k \sigma_i \, \mathbf{v}_i^\mathsf{T} \mathbf{v}_j \mathbf{u}_i \mathbf{v}_j^\mathsf{T} \;=\; \sum_{i=1}^k \sigma_i \, \mathbf{u}_i \mathbf{v}_i^\mathsf{T},$$

since the terms in the double summation are non-zero only if $1 \leq i = j \leq k$. This shows that $\hat{A} = AP_k$. Moreover,

$$\sum_{i=1}^m \sum_{j=1}^n (A_{i,j} - \hat{A}_{i,j})^2 \;=\; \sum_{i=1}^m \|\mathbf{a}_i - P_k \mathbf{a}_i\|^2 \;=\; \mathrm{cost}(S_k; \mathbf{a}_1, \ldots, \mathbf{a}_m),$$

where $\mathbf{a}_i^\mathsf{T}$ is the ith row of A, and cost is as we defined in Section 1. By Proposition 1, we have $\mathrm{cost}(S_k; \mathbf{a}_1, \ldots, \mathbf{a}_m) = \sum_{j=k+1}^r \sigma_j^2$. Moreover, by Theorem 1, we have

$$\mathrm{cost}(S_k; \mathbf{a}_1, \ldots, \mathbf{a}_m) \;\leq\; \mathrm{cost}(\mathsf{CS}(B^\mathsf{T}); \mathbf{a}_1, \ldots, \mathbf{a}_m),$$

since B has rank k.

Now we show that $\mathrm{cost}(\mathsf{CS}(B^\mathsf{T}); \mathbf{a}_1, \ldots, \mathbf{a}_m) \leq \sum_{i=1}^m \sum_{j=1}^n (A_{i,j} - B_{i,j})^2$. Let $\mathbf{b}_1^\mathsf{T}, \ldots, \mathbf{b}_m^\mathsf{T}$ be the rows of B, and let P be the orthoprojector for $\mathsf{CS}(B^\mathsf{T})$. Then, by the Pythagorean Theorem and the fact $P\mathbf{b}_i = \mathbf{b}_i$,

$$\begin{aligned}
\|\mathbf{a}_i - \mathbf{b}_i\|^2 &= \|P(\mathbf{a}_i - \mathbf{b}_i)\|^2 + \|(I - P)(\mathbf{a}_i - \mathbf{b}_i)\|^2 \\
&= \|P(\mathbf{a}_i - \mathbf{b}_i)\|^2 + \|\mathbf{a}_i - P\mathbf{a}_i + \mathbf{b}_i - P\mathbf{b}_i\|^2 \\
&\geq \|\mathbf{a}_i - P\mathbf{a}_i\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{cost}(\mathsf{CS}(B^\mathsf{T}); \mathbf{a}_1, \ldots, \mathbf{a}_m) &= \sum_{i=1}^m \|\mathbf{a}_i - P\mathbf{a}_i\|^2 \\
&\leq \sum_{i=1}^m \|\mathbf{a}_i - \mathbf{b}_i\|^2 \;=\; \sum_{i=1}^m \sum_{j=1}^m (A_{i,j} - B_{i,j})^2. \quad \square
\end{aligned}$$