

13.11 Low-Rank Approximation

The compact singular value decomposition writes a matrix as a sum of rank-one pieces arranged in decreasing order of size. A sum can be cut short. This section shows that cutting it short after k terms produces the **best** approximation to $\mathbf{A}$ by a matrix of rank k , measured in the Frobenius norm, and that the error it leaves is a quantity we can write down before doing any work. One theorem answers three questions at once: how to compress a matrix, which k -dimensional subspace best fits a cloud of data, and how far a given matrix is from being singular.

Throughout, $F = \mathbb{R}$ or $F = \mathbb{C}$, $\mathbf{A} \in M_{m \times n}(F)$, $p = \min(m, n)$, $r = \text{rank } \mathbf{A}$, and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$$

are the singular values of $\mathbf{A}$ (Definition 13.9.1), so that $\sigma_i > 0$ exactly for $i \leq r$. We fix a singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$ (Theorem 13.9.1) and write $\mathbf{u}_1, \dots, \mathbf{u}_m$ for the columns of $\mathbf{U}$ and $\mathbf{v}_1, \dots, \mathbf{v}_n$ for the columns of $\mathbf{V}$.

13.11.1 Cutting the sum short

Multiplying out $\mathbf{U}\Sigma\mathbf{V}^*$ column by column gives the rank-one expansion

$$\mathbf{A} = \sum_{i=1}^p \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

the second equality because the terms with $i > r$ have $\sigma_i = 0$; this is the compact form Theorem 13.9.3. Each summand is a matrix of rank one, and the coefficients decrease. The natural thing to do with such a sum is to keep the large terms and discard the small ones.

Keep the first k rank-one pieces and throw the rest away.

Definition 13.11.1. Truncated singular value decomposition

For $0 \leq k \leq p$, the **rank- k truncation** of $\mathbf{A}$ relative to the chosen singular value decomposition $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$ (Theorem 13.9.1) is

$$\mathbf{A}_k := \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

with $\mathbf{A}_0 = \mathbf{O}$; for $k \geq p$ we set $\mathbf{A}_k = \mathbf{A}$.

Two clauses need comment. The subscript counts **terms kept**, not rows and columns: $\mathbf{A}_k$ is the same size as $\mathbf{A}$, not a submatrix of it. (Section 1 used $\mathbf{A}_k$ for the top-left $k \times k$ corner; that meaning is gone for the rest of this section.) And the definition says “relative to the chosen decomposition”, because Theorem 13.9.2 leaves the unitary factors some freedom, and different choices can give different truncations. Everything proved below holds for **every** choice; the warning at the end of the section shows that the dependence is real.

The rank of $\mathbf{A}_k$ is exactly $\min(k, r)$: the vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ are orthonormal, hence independent, so the columns of $\mathbf{A}_k$ span $\text{span}(\mathbf{u}_1, \dots, \mathbf{u}_{\min(k, r)})$ once the vanishing terms are dropped.

Example. Truncating a diagonal matrix

Let $\mathbf{A} = \text{diag}(5, 3, 1) \in M_3(\mathbb{R})$. Write down $\mathbf{A}_1$ and $\mathbf{A}_2$.

Solution. A diagonal matrix with positive decreasing entries is already in the form $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ with $\mathbf{U} = \mathbf{V} = \mathbf{I}_3$, so $\sigma_i = 5, 3, 1$ and $\mathbf{u}_i = \mathbf{v}_i = \mathbf{e}_i$. Hence

$$\mathbf{A}_1 = 5\mathbf{e}_1\mathbf{e}_1^* = \text{diag}(5, 0, 0), \quad \mathbf{A}_2 = \text{diag}(5, 3, 0).$$

Truncation of a diagonal matrix is just deleting its smallest diagonal entries. Everything in this section is this statement, conjugated by two unitary matrices.

The first thing to compute is how much was thrown away.

Lemma 13.11.1. The Frobenius norm counts singular values

Let $\mathbf{A} \in M_{m \times n}(F)$ have singular values $\sigma_1 \geq \dots \geq \sigma_p$. Then (the first identity is the one Section 9 already obtained, recorded here with a label because the proof of the main theorem uses it repeatedly)

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^p \sigma_i^2, \quad \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$$

for every $0 \leq k \leq p$.

Proof. The Frobenius norm does not see unitary factors (Lemma 11.7.5), so from $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$,

$$\|\mathbf{A}\|_F = \|\mathbf{\Sigma}\mathbf{V}^*\|_F = \|\mathbf{\Sigma}\|_F.$$

The only nonzero entries of $\mathbf{\Sigma}$ are $\sigma_1, \dots, \sigma_p$ on the diagonal, so $\|\mathbf{\Sigma}\|_F^2 = \sum_i \sigma_i^2$.

For the second identity, subtract the first k terms of the rank-one expansion:

$$\mathbf{A} - \mathbf{A}_k = \sum_{i>k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{U}\mathbf{\Sigma}^{(k)}\mathbf{V}^*,$$

where $\mathbf{\Sigma}^{(k)}$ is $\mathbf{\Sigma}$ with its first k diagonal entries replaced by 0. Applying the first identity to this matrix — the display is a singular value decomposition of it in the sense of Theorem 13.9.1, up to reordering the diagonal — gives $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$. ■

So the truncation error is decided in advance by the singular values, and it is small exactly when the discarded σ_i are small. The content of the section is that no other matrix of rank k does better.

Quick check. A matrix has singular values 5, 4, 3. What is the smallest k for which the relative error $\|\mathbf{A} - \mathbf{A}_k\|_F / \|\mathbf{A}\|_F$ is at most $\frac{1}{2}$?

Solution. By Lemma 13.11.1, $\|\mathbf{A}\|_F^2 = 25 + 16 + 9 = 50$. For $k = 1$ the error squared is $16 + 9 = 25$, so the relative error is $\sqrt{25/50} = 1/\sqrt{2} \approx 0.71$, too big. For $k = 2$ the error squared is 9, and $\sqrt{9/50} \approx 0.42 \leq \frac{1}{2}$. So $k = 2$. Note that the test is on the *squares*: halving the error means keeping three quarters of $\sum \sigma_i^2$.

13.11.2 How much a subspace can hold

The proof of the theorem needs one inequality, and it is the only place where any work happens. The question it answers is: if we are allowed to keep only an s -dimensional subspace of F^m , how much of $\mathbf{A}$ can we capture? The answer is $\sigma_1^2 + \cdots + \sigma_s^2$, and the best subspace is the span of the leading left singular vectors.

Lemma 13.11.2. An orthonormal list captures at most the leading singular values

Let $\mathbf{A} \in M_{m \times n}(F)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_p$, extended by $\sigma_i = 0$ for $p < i \leq m$. Let $\mathbf{q}_1, \dots, \mathbf{q}_s$ be an orthonormal list in F^m , so that $s \leq m$. Then

$$\sum_{j=1}^s \|\mathbf{A}^* \mathbf{q}_j\|^2 \leq \sum_{i=1}^s \sigma_i^2.$$

Idea. Write each $\mathbf{q}_j$ in the orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_m)$. The quantity $\|\mathbf{A}^* \mathbf{q}_j\|^2$ then becomes $\sum_i \sigma_i^2 |c_{ji}|^2$, so the total is $\sum_i \sigma_i^2 t_i$ with weights $t_i = \sum_j |c_{ji}|^2$. Two facts pin the weights down: each $t_i \leq 1$, because the $\mathbf{q}_j$ are orthonormal and Bessel's inequality applies to $\mathbf{u}_i$; and $\sum_i t_i = s$, because each $\mathbf{q}_j$ is a unit vector. A weighted sum of a decreasing sequence with weights in $[0, 1]$ that sum to s is largest when the first s weights are 1 — which is the last display of the proof.

Proof. If $s = 0$ both sides are empty sums, so assume $s \geq 1$. Fix j and put $c_{ji} = \langle \mathbf{q}_j, \mathbf{u}_i \rangle$ for $1 \leq i \leq m$, so that the i -th entry of $\mathbf{U}^* \mathbf{q}_j$ is c_{ji} . Since $\mathbf{A}^* = \mathbf{V} \mathbf{\Sigma}^* \mathbf{U}^*$ and $\mathbf{V}$ is unitary, hence an isometry (Definition 11.7.2, Theorem 11.7.3),

$$\|\mathbf{A}^* \mathbf{q}_j\|^2 = \|\mathbf{\Sigma}^* \mathbf{U}^* \mathbf{q}_j\|^2 = \sum_{i=1}^m \sigma_i^2 |c_{ji}|^2,$$

because $\mathbf{\Sigma}^*$ multiplies the i -th entry by σ_i and discards the entries beyond the p -th, which is the same as using $\sigma_i = 0$ there. Summing over j and exchanging the two finite sums,

$$\sum_{j=1}^s \|\mathbf{A}^* \mathbf{q}_j\|^2 = \sum_{i=1}^m \sigma_i^2 t_i, \quad t_i := \sum_{j=1}^s |c_{ji}|^2.$$

Now bound the weights. Since $|c_{ji}| = |\langle \mathbf{u}_i, \mathbf{q}_j \rangle|$ and $(\mathbf{q}_1, \dots, \mathbf{q}_s)$ is an orthonormal list, Bessel's inequality (Theorem 11.9.1) applied to the vector $\mathbf{u}_i$ gives

$$t_i = \sum_{j=1}^s |\langle \mathbf{u}_i, \mathbf{q}_j \rangle|^2 \leq \|\mathbf{u}_i\|^2 = 1.$$

Since $(\mathbf{u}_1, \dots, \mathbf{u}_m)$ is an orthonormal **basis** of F^m , Parseval's identity (Theorem 11.9.2) gives $\sum_{i=1}^m |c_{ji}|^2 = \|\mathbf{q}_j\|^2 = 1$ for each j , and therefore

$$\sum_{i=1}^m t_i = \sum_{j=1}^s \sum_{i=1}^m |c_{ji}|^2 = s.$$

Finally, using $\sigma_i^2 \geq \sigma_s^2$ together with $t_i - 1 \leq 0$ for $i \leq s$, and $\sigma_i^2 \leq \sigma_s^2$ together with $t_i \geq 0$ for $i > s$,

$$\begin{aligned} \sum_{i=1}^m \sigma_i^2 t_i - \sum_{i=1}^s \sigma_i^2 &= \sum_{i \leq s} \sigma_i^2 (t_i - 1) + \sum_{i > s} \sigma_i^2 t_i \\ &\leq \sigma_s^2 \left(\sum_{i \leq s} (t_i - 1) + \sum_{i > s} t_i \right) \\ &= \sigma_s^2 \left(\sum_{i=1}^m t_i - s \right) = 0. \end{aligned}$$

This is the stated inequality. ■

Equality is attainable: take $\mathbf{q}_j = \mathbf{u}_j$ for $j \leq s$. Then $\mathbf{A}^* \mathbf{u}_j = \sigma_j \mathbf{v}_j$ by the rank-one expansion, so $\|\mathbf{A}^* \mathbf{u}_j\|^2 = \sigma_j^2$ and the two sides agree. The lemma is the extreme case of a family of statements about sums of the largest eigenvalues of a Hermitian matrix; Chapter 17 develops the family through the Courant–Fischer min–max theorem, of which the inequality above is the crudest and most useful consequence.

13.11.3 The Eckart–Young theorem

Now the theorem. Read it as a statement about a minimization problem: over the set of matrices of rank at most k — which is not a subspace, so no earlier optimization theorem applies to it directly — the minimum of $\|\mathbf{A} - \mathbf{B}\|_F$ exists, is attained at a matrix we can write down, and has a value we already computed.

Theorem 13.11.3. Eckart–Young Theorem, Frobenius case

Let $\mathbf{A} \in M_{m \times n}(F)$ with singular values $\sigma_1 \geq \dots \geq \sigma_p$, and let $0 \leq k \leq p$. Then for **every** $\mathbf{B} \in M_{m \times n}(F)$ with $\text{rank } \mathbf{B} \leq k$,

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \left(\sum_{i > k} \sigma_i^2 \right)^{1/2},$$

and equality holds for $\mathbf{B} = \mathbf{A}_k$. In words: the truncation $\mathbf{A}_k$ minimizes $\|\mathbf{A} - \mathbf{B}\|_F$ over all matrices of rank at most k , and the minimum value is $(\sum_{i > k} \sigma_i^2)^{1/2}$.

Idea. The difficulty is that $\mathbf{B}$ is an arbitrary matrix of rank $\leq k$, and there is no formula for such a thing. The rescue is to stop looking at $\mathbf{B}$ and look only at its column space S , a subspace of dimension $s \leq k$. ① Every column of $\mathbf{B}$ lies in S , so column by column the Best Approximation Theorem says $\mathbf{B}$ cannot beat $P_S \mathbf{A}$, the matrix obtained by projecting every column of $\mathbf{A}$ onto S . So it suffices to minimize over subspaces. ② For a subspace, Pythagoras splits the error into “all of $\mathbf{A}$ ” minus “the part S captures”, and the captured part is exactly

the quantity [Lemma 13.11.2](#) bounds. ③ The bound is $\sigma_1^2 + \cdots + \sigma_k^2$, and subtracting it from $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ leaves $\sum_{i>k} \sigma_i^2$, which [Lemma 13.11.1](#) says is exactly what $\mathbf{A}_k$ achieves.

Proof. Step 1: reduce to a projection. Let $\mathbf{B}$ have rank $\leq k$ and put $S = \text{col}(\mathbf{B})$, so $s := \dim S = \text{rank } \mathbf{B} \leq k$. Write $\mathbf{a}_1, \dots, \mathbf{a}_n$ and $\mathbf{b}_1, \dots, \mathbf{b}_n$ for the columns of $\mathbf{A}$ and of $\mathbf{B}$. By the definition of the Frobenius norm, summing squares of entries column by column,

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{j=1}^n \|\mathbf{a}_j - \mathbf{b}_j\|^2.$$

Each $\mathbf{b}_j$ lies in S , so Theorem 11.3.5 gives $\|\mathbf{a}_j - \mathbf{b}_j\| \geq \|\mathbf{a}_j - P_S \mathbf{a}_j\|$ for every j , where P_S is the orthogonal projection onto S (Definition 11.3.2), identified throughout with its matrix in the standard basis of F^m . Hence

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \geq \sum_{j=1}^n \|\mathbf{a}_j - P_S \mathbf{a}_j\|^2 = \|\mathbf{A} - P_S \mathbf{A}\|_F^2,$$

the last equality because the j -th column of $P_S \mathbf{A}$ is $P_S \mathbf{a}_j$.

Step 2: what a subspace captures. Choose an orthonormal basis $(\mathbf{q}_1, \dots, \mathbf{q}_s)$ of S . For each j , the vectors $\mathbf{a}_j - P_S \mathbf{a}_j$ and $P_S \mathbf{a}_j$ are orthogonal, so Theorem 11.1.3 gives $\|\mathbf{a}_j - P_S \mathbf{a}_j\|^2 = \|\mathbf{a}_j\|^2 - \|P_S \mathbf{a}_j\|^2$, and summing over j ,

$$\|\mathbf{A} - P_S \mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2 - \|P_S \mathbf{A}\|_F^2.$$

By Theorem 11.3.3 (a), $P_S \mathbf{a}_j = \sum_{l \leq s} \langle \mathbf{a}_j, \mathbf{q}_l \rangle \mathbf{q}_l$, whose summands are pairwise orthogonal, so Theorem 11.1.3 gives $\|P_S \mathbf{a}_j\|^2 = \sum_{l=1}^s |\langle \mathbf{a}_j, \mathbf{q}_l \rangle|^2$. Exchanging the sums,

$$\|P_S \mathbf{A}\|_F^2 = \sum_{l=1}^s \sum_{j=1}^n |\langle \mathbf{a}_j, \mathbf{q}_l \rangle|^2 = \sum_{l=1}^s \|\mathbf{A}^* \mathbf{q}_l\|^2,$$

where the last equality holds because $\langle \mathbf{a}_j, \mathbf{q}_l \rangle = \mathbf{q}_l^* \mathbf{a}_j$ is the j -th entry of the row vector $\mathbf{q}_l^* \mathbf{A}$, whose entries are the conjugates of those of $\mathbf{A}^* \mathbf{q}_l$.

Step 3: apply the bound. By [Lemma 13.11.2](#) and $s \leq k$, and since all $\sigma_i^2 \geq 0$,

$$\|P_S \mathbf{A}\|_F^2 \leq \sum_{i=1}^s \sigma_i^2 \leq \sum_{i=1}^k \sigma_i^2.$$

Combining the three steps with $\|\mathbf{A}\|_F^2 = \sum_i \sigma_i^2$ ([Lemma 13.11.1](#)),

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \geq \|\mathbf{A}\|_F^2 - \sum_{i=1}^k \sigma_i^2 = \sum_{i>k} \sigma_i^2.$$

Equality. The matrix $\mathbf{A}_k$ has rank $\min(k, r) \leq k$, and $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$ by [Lemma 13.11.1](#). So the lower bound is attained, and it is the minimum. This proves the theorem. ■

Three remarks are worth extracting. First, the theorem is a genuine minimum, not an infimum: the minimizing matrix is exhibited. Second, the minimum value depends on $\mathbf{A}$ only through its singular values, so it is unchanged if $\mathbf{A}$ is replaced by $\mathbf{U}'\mathbf{A}\mathbf{V}'$ with $\mathbf{U}', \mathbf{V}'$ unitary. Third, the case $k = r - 1$ reads: the Frobenius distance from $\mathbf{A}$ to the matrices of rank $< r$ is σ_r , the smallest nonzero singular value. Exercise C1 turns that into a statement about how far an invertible matrix is from being singular.

Remark. The same matrix $\mathbf{A}_k$ is also optimal in the **spectral norm** $\|\cdot\|_2$, where the minimum value is σ_{k+1} rather than $(\sum_{i>k} \sigma_i^2)^{1/2}$. We cannot state that here, because the spectral norm is defined only in Chapter 16. In fact both statements are instances of one theorem: $\mathbf{A}_k$ is a best rank- k approximation in **every** unitarily invariant norm. Chapter 21 proves that, and the proof is genuinely different from the one above — it needs inequalities between the singular values of $\mathbf{A}$ and of $\mathbf{A} - \mathbf{B}$, not a projection argument.

13.11.4 A worked example

Example. The best rank-one approximation of a 3×2 matrix

Let

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \in M_{3 \times 2}(\mathbb{R}).$$

Find the best rank-one approximation of $\mathbf{A}$ in the Frobenius norm, and compute the error in two ways.

Solution. Singular values and vectors. Since $\mathbf{A}$ has more rows than columns, work with the 2×2 matrix

$$\mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

whose eigenvalues are 3 and 1 with orthonormal eigenvectors

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, 1), \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}(1, -1).$$

Hence $\sigma_1 = \sqrt{3}$ and $\sigma_2 = 1$, and the left singular vectors are $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i/\sigma_i$:

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}}(1, 1, 2), \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1, 0).$$

(Both are unit vectors, as they must be: $1 + 1 + 4 = 6$ and $1 + 1 = 2$.)

The approximation. By [Theorem 13.11.3](#) the answer is $\mathbf{A}_1 = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top$. The scalar in front is $\sqrt{3}/(\sqrt{6}\sqrt{2}) = \sqrt{3}/\sqrt{12} = \frac{1}{2}$, so

$$\mathbf{A}_1 = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{pmatrix}.$$

It has rank one, its two columns being equal and nonzero.

The error, twice. Directly,

$$\mathbf{A} - \mathbf{A}_1 = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \|\mathbf{A} - \mathbf{A}_1\|_F^2 = 4 \cdot \frac{1}{4} = 1.$$

From the theorem, the error squared is $\sigma_2^2 = 1$. The two agree. As a further check, $\|\mathbf{A}\|_F^2 = 1 + 1 + 1 + 1 = 4$ and $\sigma_1^2 + \sigma_2^2 = 3 + 1 = 4$, as [Lemma 13.11.1](#) requires; the relative error of the rank-one approximation is $\sqrt{1/4} = \frac{1}{2}$.

13.11.5 Best-fitting subspaces

Step 1 of the proof did more than its job. It showed that the optimal $\mathbf{B}$ may be taken of the form $P_S \mathbf{A}$, and Steps 2 and 3 identified the best S . That identification is a theorem in its own right, and it is the statement people mean when they speak of fitting a low-dimensional structure to data.

Corollary 13.11.4. The best-fitting subspace

Let $\mathbf{a}_1, \dots, \mathbf{a}_n \in F^m$ be the columns of $\mathbf{A}$, and let $1 \leq k \leq p$. Among all subspaces $S \subseteq F^m$ with $\dim S \leq k$, the total squared distance

$$D(S) := \sum_{j=1}^n d(\mathbf{a}_j, S)^2$$

is smallest for $S = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$, and its smallest value is $\sum_{i>k} \sigma_i^2$.

Proof. By Definition 11.3.3, $d(\mathbf{a}_j, S) = \|\mathbf{a}_j - P_S \mathbf{a}_j\|$, so $D(S) = \|\mathbf{A} - P_S \mathbf{A}\|_F^2$ as in Step 1 of [Theorem 13.11.3](#). Since $\text{rank}(P_S \mathbf{A}) \leq \dim S \leq k$, the theorem gives $D(S) \geq \sum_{i>k} \sigma_i^2$. For $S_0 = \text{span}(\mathbf{u}_1, \dots, \mathbf{u}_k)$ the columns $\mathbf{u}_1, \dots, \mathbf{u}_k$ are orthonormal, so Theorem 11.3.3 (a) gives

$$P_{S_0} \mathbf{A} = \sum_{i=1}^k \mathbf{u}_i \mathbf{u}_i^* \mathbf{A} = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^* = \mathbf{A}_k,$$

using $\mathbf{u}_i^* \mathbf{A} = \sigma_i \mathbf{v}_i^*$, which is the rank-one expansion read one row at a time. Hence $D(S_0) = \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i>k} \sigma_i^2$. ■

What this is used for, honestly. Two applications are within reach of what we have proved, and a third is not.

The first is **compression**. Storing $\mathbf{A}$ costs mn numbers; storing $\mathbf{A}_k$ as $\sigma_1, \mathbf{u}_1, \mathbf{v}_1, \dots, \sigma_k, \mathbf{u}_k, \mathbf{v}_k$ costs $k(m + n + 1)$. When k is much smaller than m and n that is a large saving, and [Theorem 13.11.3](#) says no other rank- k object stored the same way is closer to $\mathbf{A}$ in the Frobenius norm. The relative error is $(\sum_{i>k} \sigma_i^2 / \sum_i \sigma_i^2)^{1/2}$, computable before choosing k . Whether that error is acceptable is not a mathematical question.

The second is **principal components**. Take n data points in $\mathbb{R}^m$ and place them as the

columns of $\mathbf{A}$. Then [Corollary 13.11.4](#) says: the k -dimensional subspace through the origin that minimizes the total squared distance to the data is spanned by $\mathbf{u}_1, \dots, \mathbf{u}_k$, the leading left singular vectors. These directions are the **principal components** of the data, and the numbers σ_i^2 measure how much of $\sum_j \|\mathbf{a}_j\|^2$ each one accounts for. That is a theorem of linear algebra, and it is the whole of what this book has proved.

What is *not* proved here is the statistics. The usual procedure subtracts the mean of the data first, so that $\frac{1}{n}\mathbf{A}\mathbf{A}^\top$ becomes a sample covariance matrix, and then interprets $\sigma_i^2 / \sum_l \sigma_l^2$ as a fraction of variance explained. Those steps need a probabilistic model, a reason to prefer squared error, and an argument that the sample answer says something about a larger population. Chapter 24 §10 supplies the first two — the model for the finite population of the data themselves, where expectation is an average and no measure theory is needed, and three reasons to prefer squared error — and says plainly that the third belongs to probability and statistics, which this book does not develop. It also says why a singular value decomposition is never computed by forming $\mathbf{A}^*\mathbf{A}$ and finding its eigenvalues, the route taken in every hand computation in this chapter; the algorithm that replaces it is assembled nowhere in this book.

Warning. **The truncation is optimal for this norm, and low rank does not mean small.** Two mistakes to avoid.

Other measures of error give other answers. Let $\mathbf{A} = \text{diag}(2, 1)$, so $\sigma_1 = 2$, $\sigma_2 = 1$, and $\mathbf{A}_1 = \text{diag}(2, 0)$, whose error matrix $\text{diag}(0, 1)$ has a largest entry of modulus 1. The rank-one matrix

$$\mathbf{B} = \frac{1}{2} \begin{pmatrix} 3 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix}, \quad \mathbf{A} - \mathbf{B} = \frac{1}{2} \begin{pmatrix} 1 & -\sqrt{3} \\ -\sqrt{3} & 1 \end{pmatrix},$$

has $\det \mathbf{B} = \frac{3}{4} - \frac{3}{4} = 0$, so $\text{rank } \mathbf{B} = 1$, and every entry of $\mathbf{A} - \mathbf{B}$ has modulus at most $\sqrt{3}/2 \approx 0.87 < 1$. So $\mathbf{B}$ beats $\mathbf{A}_1$ if the error is measured entrywise. It loses in the Frobenius norm, as it must: $\|\mathbf{A} - \mathbf{B}\|_F^2 = 2 > 1 = \|\mathbf{A} - \mathbf{A}_1\|_F^2$.

Small rank and small size are unrelated. The $n \times n$ all-ones matrix $\mathbf{J}$ has rank 1 and $\|\mathbf{J}\|_F = n$, while $\varepsilon \mathbf{I}_n$ has rank n and $\|\varepsilon \mathbf{I}_n\|_F = \varepsilon \sqrt{n}$, as small as we like. “Approximating by a matrix of low rank” is a statement about structure, never about magnitude.

13.11.6 Exercises

13.11.6.1 A. Check your understanding

Exercise. A1

- State the Eckart–Young Theorem, including the value of the minimum.
- Express $\|\mathbf{A}\|_F$ in terms of the singular values of $\mathbf{A}$.
- Let $\text{rank } \mathbf{A} = r$. What is $\mathbf{A}_k$ for $k \geq r$, and what is the minimum of $\|\mathbf{A} - \mathbf{B}\|_F$ over $\text{rank } \mathbf{B} \leq k$ in that case?
- True or false: the truncation $\mathbf{A}_k$ is the *unique* matrix of rank at most k nearest to $\mathbf{A}$ in the Frobenius norm. Justify your answer.
- A matrix of rank 1 is close to $\mathbf{O}$ in the Frobenius norm. True or false? Justify your answer.

Solution.

- (a) For $\mathbf{A} \in M_{m \times n}(F)$ with singular values $\sigma_1 \geq \dots \geq \sigma_p$ and $0 \leq k \leq p$: every $\mathbf{B}$ with $\text{rank } \mathbf{B} \leq k$ satisfies $\|\mathbf{A} - \mathbf{B}\|_F \geq (\sum_{i>k} \sigma_i^2)^{1/2}$, with equality for $\mathbf{B} = \mathbf{A}_k = \sum_{i \leq k} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$.
- (b) $\|\mathbf{A}\|_F = (\sum_{i=1}^p \sigma_i^2)^{1/2}$ (Lemma 13.11.1).
- (c) $\mathbf{A}_k = \mathbf{A}$, since the terms with $i > r$ are zero; the minimum is 0, attained at $\mathbf{B} = \mathbf{A}$ itself, which has rank $r \leq k$.
- (d) False in general. Take $\mathbf{A} = \mathbf{I}_2$, so $\sigma_1 = \sigma_2 = 1$. Both $\mathbf{B}_1 = \text{diag}(1, 0)$ and $\mathbf{B}_2 = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ have rank 1, and $\|\mathbf{I}_2 - \mathbf{B}_1\|_F^2 = 1 = \|\mathbf{I}_2 - \mathbf{B}_2\|_F^2$, which is the minimum σ_2^2 . (Both are truncations of $\mathbf{I}_2$, for two different singular value decompositions.)
- (e) False. Rank is unrelated to size: the all-ones matrix $\mathbf{J} \in M_n(\mathbb{R})$ has rank 1 and $\|\mathbf{J}\|_F = n$.

13.11.6.2 B. Practice

Exercise. B1: Errors from the singular values alone

A matrix $\mathbf{A}$ has singular values 6, 3, 2, 1.

- (a) Compute $\|\mathbf{A}\|_F$ and the minimum of $\|\mathbf{A} - \mathbf{B}\|_F$ over matrices $\mathbf{B}$ of rank at most k , for $k = 1, 2, 3$.
- (b) Find the smallest k for which the relative error $\|\mathbf{A} - \mathbf{A}_k\|_F / \|\mathbf{A}\|_F$ is at most 0.3.

Solution.

- (a) $\|\mathbf{A}\|_F^2 = 36 + 9 + 4 + 1 = 50$, so $\|\mathbf{A}\|_F = \sqrt{50} = 5\sqrt{2}$. By Theorem 13.11.3 the minima are

$$\sqrt{9 + 4 + 1} = \sqrt{14}, \quad \sqrt{4 + 1} = \sqrt{5}, \quad \sqrt{1} = 1$$

for $k = 1, 2, 3$.

- (b) The condition is $\sum_{i>k} \sigma_i^2 \leq 0.09 \cdot 50 = 4.5$. For $k = 2$ the left side is $5 > 4.5$; for $k = 3$ it is $1 \leq 4.5$. So $k = 3$.

Exercise. B2: A best rank-one approximation

Let $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \in M_2(\mathbb{R})$. Find the singular values of $\mathbf{A}$, the best rank-one approximation $\mathbf{A}_1$ in the Frobenius norm, and the error, computed both directly and from the theorem.

Solution. $\mathbf{A}$ is symmetric with eigenvalues 4 and 2 (trace 6, determinant 8) and orthonormal eigenvectors $\mathbf{q}_1 = \frac{1}{\sqrt{2}}(1, 1)$, $\mathbf{q}_2 = \frac{1}{\sqrt{2}}(1, -1)$. Both eigenvalues are positive, so $\mathbf{A} \succ \mathbf{O}$, and $\mathbf{A}^\top \mathbf{A} = \mathbf{A}^2$ has eigenvalues 16 and 4; hence $\sigma_1 = 4$, $\sigma_2 = 2$, with $\mathbf{u}_i = \mathbf{v}_i = \mathbf{q}_i$.

Therefore

$$\mathbf{A}_1 = 4\mathbf{q}_1\mathbf{q}_1^\top = 4 \cdot \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

which has rank one. Directly,

$$\mathbf{A} - \mathbf{A}_1 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \|\mathbf{A} - \mathbf{A}_1\|_F^2 = 4.$$

From [Theorem 13.11.3](#) the error squared is $\sigma_2^2 = 4$. The two agree, and $\|\mathbf{A}\|_F^2 = 9+1+1+9 = 20 = 16 + 4$ checks [Lemma 13.11.1](#).

Exercise. B3: Pythagoras for the truncation

Let $\mathbf{A} \in M_{m \times n}(F)$ and $0 \leq k \leq p$. Prove that

$$\|\mathbf{A}\|_F^2 = \|\mathbf{A}_k\|_F^2 + \|\mathbf{A} - \mathbf{A}_k\|_F^2,$$

and deduce that $k \mapsto \|\mathbf{A} - \mathbf{A}_k\|_F$ is nonincreasing.

Solution. By [Lemma 13.11.1](#) applied to $\mathbf{A}$ and to $\mathbf{A} - \mathbf{A}_k$, and by the same lemma applied to $\mathbf{A}_k$, whose singular values are $\sigma_1, \dots, \sigma_k$ followed by zeros (it equals $\mathbf{U}\mathbf{\Sigma}'\mathbf{V}^*$ with $\mathbf{\Sigma}'$ the matrix keeping only the first k diagonal entries of $\mathbf{\Sigma}$),

$$\|\mathbf{A}_k\|_F^2 + \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i \leq k} \sigma_i^2 + \sum_{i > k} \sigma_i^2 = \|\mathbf{A}\|_F^2.$$

Since $\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i > k} \sigma_i^2$ and the σ_i^2 are nonnegative, increasing k by one removes a nonnegative term, so the sequence is nonincreasing. This proves both claims.

13.11.6.3 C. Going deeper

Exercise. C1: How far is a matrix from being singular?

Let $\mathbf{A} \in M_n(F)$ be invertible, with singular values $\sigma_1 \geq \dots \geq \sigma_n > 0$.

- Prove that $\|\mathbf{A} - \mathbf{B}\|_F \geq \sigma_n$ for every singular $\mathbf{B} \in M_n(F)$.
- Exhibit a singular $\mathbf{B}$ with $\|\mathbf{A} - \mathbf{B}\|_F = \sigma_n$.
- Deduce that the distance from $\mathbf{A}$ to the set of singular matrices, measured in the Frobenius norm, is exactly σ_n , and illustrate this with $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$.

Solution.

- A singular $\mathbf{B} \in M_n(F)$ has $\text{rank } \mathbf{B} \leq n - 1$ ([Theorem 7.5.3 \(d\)](#)). By [Theorem 13.11.3](#) with $k = n - 1$,

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \left(\sum_{i > n-1} \sigma_i^2 \right)^{1/2} = \sigma_n.$$

- (b) Take $\mathbf{B} = \mathbf{A}_{n-1} = \sum_{i \leq n-1} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$. Its rank is $n - 1 < n$, so it is singular, and $\|\mathbf{A} - \mathbf{B}\|_F = \sigma_n$ by [Lemma 13.11.1](#).
- (c) By (a) the distance is at least σ_n and by (b) it is at most σ_n , so it equals σ_n ; in particular the minimum is attained. For the given $\mathbf{A}$, Exercise B2 found $\sigma_2 = 2$ and $\mathbf{A}_1 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$, which is singular and at Frobenius distance 2. So no singular matrix is closer to $\mathbf{A}$ than 2. The smallest singular value is therefore a quantitative version of invertibility: $\mathbf{A}$ is invertible exactly when $\sigma_n > 0$, and it is *far* from singular exactly when σ_n is large.

Exercise. C2: The best line through a cloud of points

The four points $(1, 1), (2, 2), (1, -1)$ and $(0, 0)$ of $\mathbb{R}^2$ are to be fitted by a line through the origin, minimizing the total squared distance from the points to the line.

- (a) Write the points as the columns of a matrix $\mathbf{A}$ and compute $\mathbf{A}\mathbf{A}^\top$ and the singular values of $\mathbf{A}$.
- (b) Use [Corollary 13.11.4](#) to find the best line and the minimum total squared distance.
- (c) Verify the answer by computing the four distances directly.

Solution.

- (a) With the points as columns,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 1 & 2 & -1 & 0 \end{pmatrix}, \quad \mathbf{A}\mathbf{A}^\top = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}.$$

The eigenvalues of $\mathbf{A}\mathbf{A}^\top$ are 10 and 2 (trace 12, determinant 20), with orthonormal eigenvectors $\mathbf{u}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $\mathbf{u}_2 = \frac{1}{\sqrt{2}}(1, -1)$. The eigenvalues of $\mathbf{A}\mathbf{A}^\top$ are the squares of the singular values, since $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\Sigma\Sigma^*\mathbf{U}^*$; hence $\sigma_1 = \sqrt{10}$ and $\sigma_2 = \sqrt{2}$.

- (b) By [Corollary 13.11.4](#) with $k = 1$, the best line is $S_0 = \text{span}(\mathbf{u}_1)$, the line $y = x$, and the minimum total squared distance is $\sigma_2^2 = 2$.
- (c) The points $(1, 1), (2, 2)$ and $(0, 0)$ lie on $y = x$, contributing 0. For $(1, -1)$, the projection onto $\text{span}((1, 1))$ is $\frac{1-1}{2}(1, 1) = (0, 0)$, so the distance squared is $1 + 1 = 2$. The total is 2, as predicted. (The point $(0, 0)$ contributes nothing to any subspace through the origin; it is there as a reminder that the corollary counts all columns, including zero ones.)

Exercise. C3: Truncation respects unitaries but not sums

Let $\mathbf{A} \in M_{m \times n}(F)$, and let $\mathbf{Q} \in M_m(F)$ and $\mathbf{R} \in M_n(F)$ be unitary (orthogonal, when $F = \mathbb{R}$).

- (a) Prove that $\mathbf{Q}\mathbf{A}_k\mathbf{R}^*$ is a best rank- k approximation of $\mathbf{Q}\mathbf{A}\mathbf{R}^*$ in the Frobenius norm.
- (b) Deduce that the function $\mathbf{A} \mapsto \min\{\|\mathbf{A} - \mathbf{B}\|_F : \text{rank } \mathbf{B} \leq k\}$ is unchanged by $\mathbf{A} \mapsto \mathbf{Q}\mathbf{A}\mathbf{R}^*$.
- (c) Give $\mathbf{A}, \mathbf{B} \in M_2(\mathbb{R})$ with $(\mathbf{A} + \mathbf{B})_1 \neq \mathbf{A}_1 + \mathbf{B}_1$ for every choice of the truncations

involved. *Hint: two rank-one matrices whose sum is a multiple of the identity.*

Solution.

- (a) If $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ is a singular value decomposition (Theorem 13.9.1), then

$$\mathbf{QAR}^* = (\mathbf{QU})\mathbf{\Sigma}(\mathbf{RV})^*$$

is one as well, since a product of unitary matrices is unitary (Proposition 11.7.4 (a)) and $\mathbf{\Sigma}$ is unchanged. Its left and right singular vectors are $\mathbf{Qu}_i$ and $\mathbf{Rv}_i$, so its rank- k truncation for this decomposition is

$$\sum_{i \leq k} \sigma_i (\mathbf{Qu}_i)(\mathbf{Rv}_i)^* = \mathbf{Q} \left(\sum_{i \leq k} \sigma_i \mathbf{u}_i \mathbf{v}_i^* \right) \mathbf{R}^* = \mathbf{QA}_k \mathbf{R}^*.$$

By Theorem 13.11.3 applied to $\mathbf{QAR}^*$, this is a best rank- k approximation.

- (b) The two matrices have the same singular values, so by Theorem 13.11.3 both minima equal $(\sum_{i > k} \sigma_i^2)^{1/2}$.
- (c) Take $\mathbf{A} = \text{diag}(1, 0)$ and $\mathbf{B} = \text{diag}(0, 1)$. Each has rank one, so $\mathbf{A}_1 = \mathbf{A}$ and $\mathbf{B}_1 = \mathbf{B}$ for every singular value decomposition, and $\mathbf{A}_1 + \mathbf{B}_1 = \mathbf{I}_2$. But $\mathbf{A} + \mathbf{B} = \mathbf{I}_2$, whose rank-one truncations all have rank one, hence are never $\mathbf{I}_2$. So $(\mathbf{A} + \mathbf{B})_1 \neq \mathbf{A}_1 + \mathbf{B}_1$. Truncation is not a linear operation, which is why “keep the biggest pieces” cannot be analyzed one summand at a time.