Chapter 6

Singular Value Decomposition

In Chapter 5, we derived a number of algorithms for computing the eigenvalues and eigenvectors of matrices $A \in \mathbb{R}^{n\times n}$. Having developed this machinery, we complete our initial discussion of numerical linear algebra by deriving and making use of one final matrix factorization that exists for *any* matrix $A \in \mathbb{R}^{m\times n}$: the singular value decomposition (SVD).

6.1 Deriving the SVD

For $A \in \mathbb{R}^{m\times n}$, we can think of the function $\vec{x} \mapsto A\vec{x}$ as a map taking points in $\mathbb{R}^n$ to points in $\mathbb{R}^m$. From this perspective, we might ask what happens to the geometry of $\mathbb{R}^n$ in the process, and in particular the effect A has on lengths of and angles between vectors.

Applying our usual starting point for eigenvalue problems, we can ask the effect that A has on the lengths of vectors by examining critical points of the ratio

$$R(\vec{x}) = \frac{\|A\vec{x}\|}{\|\vec{x}\|}$$

over various values of $\vec{x}$. Scaling $\vec{x}$ does not matter, since

$$R(\alpha\vec{x}) = \frac{\|A\cdot\alpha\vec{x}\|}{\|\alpha\vec{x}\|} = \frac{|\alpha|}{|\alpha|}\cdot\frac{\|A\vec{x}\|}{\|\vec{x}\|} = \frac{\|A\vec{x}\|}{\|\vec{x}\|} = R(\vec{x}).$$

Thus, we can restrict our search to $\vec{x}$ with $\|\vec{x}\| = 1$. Furthermore, since $R(\vec{x}) \geq 0$, we can instead consider $[R(\vec{x})]^2 = \|A\vec{x}\|^2 = \vec{x}^\top A^\top A\vec{x}$. As we have shown in previous chapters, however, critical points of $\vec{x}^\top A^\top A\vec{x}$ subject to $\|\vec{x}\| = 1$ are exactly the eigenvectors $\vec{x}_i$ satisfying $A^\top A\vec{x}_i = \lambda_i\vec{x}_i$; notice $\lambda_i \geq 0$ and $\vec{x}_i \cdot \vec{x}_j = 0$ when $i \neq j$ since $A^\top A$ is symmetric and positive semidefinite.

Based on our use of the function R, the $\{\vec{x}_i\}$ basis is a reasonable one for studying the geometric effects of A. Returning to this original goal, define $\vec{y}_i \equiv A\vec{x}_i$. We can make an additional observation about $\vec{y}_i$ revealing even stronger eigenvalue structure:

$$\begin{aligned}
\lambda_i\vec{y}_i &= \lambda_i \cdot A\vec{x}_i \text{ by definition of } \vec{y}_i\\
&= A(\lambda_i\vec{x}_i)\\
&= A(A^\top A\vec{x}_i) \text{ since } \vec{x}_i \text{ is an eigenvector of } A^\top A\\
&= (AA^\top)(A\vec{x}_i) \text{ by associativity}\\
&= (AA^\top)\vec{y}_i
\end{aligned}$$

Thus, we have two cases:

1. When $\lambda_i \neq 0$, then $\vec{y}_i \neq \vec{0}$. In this case, $\vec{x}_i$ is an eigenvector of $A^\top A$ and $\vec{y}_i = A\vec{x}_i$ is a corresponding eigenvector of $AA^\top$ with $\|\vec{y}_i\| = \|A\vec{x}_i\| = \sqrt{\|A\vec{x}_i\|^2} = \sqrt{\vec{x}_i^\top A^\top A\vec{x}_i} = \sqrt{\lambda_i}\|\vec{x}_i\|$.
2. When $\lambda_i = 0$, $\vec{y}_i = \vec{0}$.

An identical proof shows that if $\vec{y}$ is an eigenvector of $AA^\top$, then $\vec{x} \equiv A^\top \vec{y}$ is either zero or an eigenvector of $A^\top A$ with the same eigenvalue.

Take k to be the number of strictly positive eigenvalues $\lambda_i > 0$ discussed above. By our construction above, we can take $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$ to be eigenvectors of $A^\top A$ and corresponding eigenvectors $\vec{y}_1, \ldots, \vec{y}_k \in \mathbb{R}^m$ of $AA^\top$ such that

$$\begin{aligned} A^\top A\vec{x}_i &= \lambda_i \vec{x}_i \\ AA^\top \vec{y}_i &= \lambda_i \vec{y}_i \end{aligned}$$

for eigenvalues $\lambda_i > 0$; here we normalize such that $\|\vec{x}_i\| = \|\vec{y}_i\| = 1$ for all i. Following traditional notation, we can define matrices $\bar{V} \in \mathbb{R}^{n\times k}$ and $\bar{U} \in \mathbb{R}^{m\times k}$ whose columns are $\vec{x}_i$'s and $\vec{y}_i$'s, resp.

We can examine the effect of these new basis matrices on A. Take $\vec{e}_i$ to be the i-th standard basis vector. Then,

$$\begin{aligned} \bar{U}^\top A\bar{V}\vec{e}_i &= \bar{U}^\top A\vec{x}_i \text{ by definition of } \bar{V} \\ &= \frac{1}{\lambda_i}\bar{U}^\top A(\lambda_i\vec{x}_i) \text{ since we assumed } \lambda_i > 0 \\ &= \frac{1}{\lambda_i}\bar{U}^\top A(A^\top A\vec{x}_i) \text{ since } \vec{x}_i \text{ is an eigenvector of } A^\top A \\ &= \frac{1}{\lambda_i}\bar{U}^\top (AA^\top)A\vec{x}_i \text{ by associativity} \\ &= \frac{1}{\sqrt{\lambda_i}}\bar{U}^\top (AA^\top)\vec{y}_i \text{ since we rescaled so that } \|\vec{y}_i\| = 1 \\ &= \sqrt{\lambda_i}\bar{U}^\top \vec{y}_i \text{ since } AA^\top\vec{y}_i = \lambda_i\vec{y}_i \\ &= \sqrt{\lambda_i}\vec{e}_i \end{aligned}$$

Take $\bar{\Sigma} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$. Then, the derivation above shows that $\bar{U}^\top A\bar{V} = \bar{\Sigma}$.

Complete the columns of $\bar{U}$ and $\bar{V}$ to $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ by adding orthonormal vectors $\vec{x}_i$ and $\vec{y}_i$ with $A^\top A\vec{x}_i = \vec{0}$ and $AA^\top\vec{y}_i = \vec{0}$, resp. In this case it is easy to show $U^\top AV\vec{e}_i = \vec{0}$ and/or $\vec{e}_i^\top U^\top AV = \vec{0}^\top$. Thus, if we take

$$\Sigma_{ij} \equiv \begin{cases} \sqrt{\lambda_i} & i = j \text{ and } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

then we can extend our previous relationship to show $U^\top AV = \Sigma$, or equivalently

$$A = U\Sigma V^\top.$$

This factorization is exactly the *singular value decomposition* (SVD) of A. The columns of U span the column space of A and are called its *left singular vectors*; the columns of V span its row space and are the *right singular vectors*. The diagonal elements σ_i of Σ are the *singular values* of A; usually they are sorted such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Both U and V are orthogonal matrices.

The SVD provides a complete geometric characterization of the action of A. Since U and V are orthogonal, they can be thought of as rotation matrices; as a diagonal matrix, Σ simply scales individual coordinates. Thus, *all* matrices $A \in \mathbb{R}^{m\times n}$ are a composition of a rotation, a scale, and a second rotation.

6.1.1 Computing the SVD

Recall that the columns of V simply are the eigenvectors of $A^\top A$, so they can be computed using techniques discussed in the previous chapter. Since $A = U\Sigma V^\top$, we know $AV = U\Sigma$. Thus, the columns of U corresponding to nonzero singular values in Σ simply are normalized columns of AV; the remaining columns satisfy $AA^\top \vec{u}_i = \vec{0}$, which can be solved using LU factorization.

This strategy is by no means the most efficient or stable approach to computing the SVD, but it works reasonably well for many applications. We will omit more specialized approaches to finding the SVD but note that that many are simple extensions of power iteration and other strategies we already have covered that operate without forming $A^\top A$ or $AA^\top$ explicitly.

6.2 Applications of the SVD

We devote the remainder of this chapter introducing many applications of the SVD. The SVD appears countless times in both the theory and practice of numerical linear linear algebra, and its importance hardly can be exaggerated.

6.2.1 Solving Linear Systems and the Pseudoinverse

In the special case where $A \in \mathbb{R}^{n\times n}$ is square and invertible, it is important to note that the SVD can be used to solve the linear problem $A\vec{x} = \vec{b}$. In particular, we have $U\Sigma V^\top \vec{x} = \vec{b}$, or

$$\vec{x} = V\Sigma^{-1}U^\top \vec{b}.$$

In this case Σ is a square diagonal matrix, so Σ^{-1} simply is the matrix whose diagonal entries are $1/\sigma_i$.

Computing the SVD is far more expensive than most of the linear solution techniques we introduced in Chapter 2, so this initial observation mostly is of theoretical interest. More generally, suppose we wish to find a least-squares solution to $A\vec{x} \approx \vec{b}$, where $A \in \mathbb{R}^{m\times n}$ is not necessarily square. From our discussion of the normal equations, we know that $\vec{x}$ must satisfy $A^\top A\vec{x} = A^\top \vec{b}$. Thus far, we mostly have disregarded the case when A is "short" or "underdetermined," that is, when A has more columns than rows. In this case the solution to the normal equations is nonunique.

To cover all three cases, we can solve an optimization problem of the following form:

$$\begin{array}{rl} \text{minimize} & \|\vec{x}\|^2 \\ \text{such that} & A^\top A\vec{x} = A^\top \vec{b} \end{array}$$

In words, this optimization asks that $\vec{x}$ satisfy the normal equations with the least possible norm.

Now, let's write $A = U\Sigma V^\top$. Then,

$$\begin{aligned} A^\top A &= (U\Sigma V^\top)^\top (U\Sigma V^\top) \\ &= V\Sigma^\top U^\top U\Sigma V^\top \text{ since } (AB)^\top = B^\top A^\top \\ &= V\Sigma^\top \Sigma V^\top \text{ since } U \text{ is orthogonal} \end{aligned}$$

Thus, asking that $A^\top A\vec{x} = A^\top \vec{b}$ is the same as asking

$$V\Sigma^\top \Sigma V^\top \vec{x} = V\Sigma U^\top \vec{b}$$
$$\text{Or equivalently, } \Sigma\vec{y} = \vec{d}$$

if we take $\vec{d} \equiv U^\top \vec{b}$ and $\vec{y} \equiv V^\top \vec{x}$. Notice that $\|\vec{y}\| = \|\vec{x}\|$ since U is orthogonal, so our optimization becomes:

$$\begin{array}{rl} \text{minimize} & \|\vec{y}\|^2 \\ \text{such that} & \Sigma\vec{y} = \vec{d} \end{array}$$

Since Σ is diagonal, however, the condition $\Sigma\vec{y} = \vec{d}$ simply states $\sigma_i y_i = d_i$; so, whenever $\sigma_i \neq 0$ we must have $y_i = {}^{d_i}/_{\sigma_i}$. When $\sigma_i = 0$, there is *no* constraint on y_i, so since we are minimizing $\|\vec{y}\|^2$ we might as well take $y_i = 0$. In other words, the solution to this optimization is $\vec{y} = \Sigma^+\vec{d}$, where $\Sigma^+ \in \mathbb{R}^{n\times m}$ has the following form:

$$\Sigma^+_{ij} \equiv \begin{cases} {}^1/_{\sigma_i} & i = j, \sigma_i \neq 0, \text{ and } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

This form in turn yields $\vec{x} = V\vec{y} = V\Sigma^+\vec{d} = V\Sigma^+U^\top\vec{b}$.

With this motivation, we make the following definition:

Definition 6.1 (Pseudoinverse). *The* pseudoinverse *of* $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ *is* $A^+ \equiv V\Sigma^+U^\top \in \mathbb{R}^{n\times m}$.

Our derivation above shows that the pseudoinverse of A enjoys the following properties:

- When A is square and invertible, $A^+ = A^{-1}$.
- When A is overdetermined, $A^+\vec{b}$ gives the least-squares solution to $A\vec{x} \approx \vec{b}$.
- When A is underdetermined, $A^+\vec{b}$ gives the least-squares solution to $A\vec{x} \approx \vec{b}$ with minimal (Euclidean) norm.

In this way, we finally are able to unify the underdetermined, fully-determined, and overdetermined cases of $A\vec{x} \approx \vec{b}$.

6.2.2 Decomposition into Outer Products and Low-Rank Approximations

If we expand out the product $A = U\Sigma V^\top$, it is easy to show that this relationship implies:

$$A = \sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top,$$

where $\ell \equiv \min\{m, n\}$, and $\vec{u}_i$ and $\vec{v}_i$ are the i-th columns of U and V, resp. Our sum only goes to $\min\{m, n\}$ since we know that the remaining columns of U or V will be zeroed out by Σ.

This expression shows that any matrix can be decomposed as the sum of *outer products* of vectors:

Definition 6.2 (Outer product)**.** *The* outer product *of $\vec{u} \in \mathbb{R}^m$ and $\vec{v} \in \mathbb{R}^n$ is the matrix $\vec{u} \otimes \vec{v} \equiv \vec{u}\vec{v}^\top \in \mathbb{R}^{m \times n}$.*

Suppose we wish to write the product $A\vec{x}$. Then, instead we could write:

$$\begin{aligned} A\vec{x} &= \left(\sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top \right) \vec{x} \\ &= \sum_{i=1}^{\ell} \sigma_i \vec{u}_i (\vec{v}_i^\top \vec{x}) \\ &= \sum_{i=1}^{\ell} \sigma_i (\vec{v}_i \cdot \vec{x}) \vec{u}_i \text{ since } \vec{x} \cdot \vec{y} = \vec{x}^\top \vec{y} \end{aligned}$$

So, applying A to $\vec{x}$ is the same as linearly combining the $\vec{u}_i$ vectors with weights $\sigma_i(\vec{v}_i \cdot \vec{x})$. This strategy for computing $A\vec{x}$ can provide considerably savings when the number of nonzero σ_i values is relatively small. Furthermore, we can ignore small values of σ_i, effectively truncating this sum to *approximate* $A\vec{x}$ with less work.

Similarly, from §6.2.1 we can write the pseudoinverse of A as:

$$A^+ = \sum_{\sigma_i \neq 0} \frac{\vec{v}_i \vec{u}_i^\top}{\sigma_i}.$$

Obviously we can apply the same trick to evaluate $A^+\vec{x}$, and in fact we can approximate $A^+\vec{x}$ by only evaluating those terms in the sum for which σ_i is relatively *small*. In practice, we compute the singular values σ_i as square roots of eigenvalues of $A^\top A$ or $AA^\top$, and methods like power iteration can be used to reveal a partial rather than full set of eigenvalues. Thus, if we are going to have to solve a number of least-squares problems $A\vec{x}_i \approx \vec{b}_i$ for different $\vec{b}_i$ and are satisfied with an approximation of $\vec{x}_i$, it can be valuable first to compute the smallest σ_i values first and use the approximation above. This strategy also avoids ever having to compute or store the full A^+ matrix and can be accurate when A has a wide range of singular values.

Returning to our original notation $A = U\Sigma V^\top$, our argument above effectively shows that a potentially useful approximation of A is $\tilde{A} \equiv U\tilde{\Sigma}V^\top$, where $\tilde{\Sigma}$ rounds small values of Σ to zero. It is easy to check that the column space of $\tilde{A}$ has dimension equal to the number of nonzero values on the diagonal of $\tilde{\Sigma}$. In fact, this approximation is not an *ad hoc* estimate but rather solves a difficult optimization problem post by the following famous theorem (stated without proof):

Theorem 6.1 (Eckart-Young, 1936)**.** *Suppose $\tilde{A}$ is obtained from $A = U\Sigma V^\top$ by truncating all but the k largest singular values σ_i of A to zero. Then $\tilde{A}$ minimizes both $\|A - \tilde{A}\|_{Fro}$ and $\|A - \tilde{A}\|_2$ subject to the constraint that the column space of $\tilde{A}$ have at most dimension k.*

6.2.3 Matrix Norms

Constructing the SVD also enables us to return to our discussion of matrix norms from §3.3.1. For example, recall that we defined the *Frobenius* norm of A as

$$\|A\|_{\text{Fro}}^2 \equiv \sum_{ij} a_{ij}^2.$$

If we write $A = U\Sigma V^\top$, we can simplify this expression:

$$\begin{aligned}
\|A\|_{\text{Fro}}^2 &= \sum_j \|A\vec{e}_j\|^2 \text{ since this product is the } j\text{-th column of } A \\
&= \sum_j \|U\Sigma V^\top \vec{e}_j\|^2, \text{ substituting the SVD} \\
&= \sum_j \vec{e}_j^\top V\Sigma^2 V^\top \vec{e}_j \text{ since } \|\vec{x}\|^2 = \vec{x}^\top\vec{x} \text{ and } U \text{ is orthogonal} \\
&= \|\Sigma V^\top\|_{\text{Fro}}^2 \text{ by the same logic} \\
&= \|V\Sigma\|_{\text{Fro}}^2 \text{ since a matrix and its transpose have the same Frobenius norm} \\
&= \sum_j \|V\Sigma\vec{e}_j\|^2 = \sum_j \sigma_j^2 \|V\vec{e}_j\|^2 \text{ by diagonality of } \Sigma \\
&= \sum_j \sigma_j^2 \text{ since } V \text{ is orthogonal}
\end{aligned}$$

Thus,the Frobenius norm of $A \in \mathbb{R}^{m\times n}$ is the sum of the squares of its singular values.

This result is of theoretical interest, but practically speaking the basic definition of the Frobenius norm is already straightforward to evaluate. More interestingly, recall that the induced two-norm of A is given by

$$\|A\|_2^2 = \max\{\lambda : \text{there exists } \vec{x} \in \mathbb{R}^n \text{ with } A^\top A\vec{x} = \lambda\vec{x}\}.$$

Now that we have studied eigenvalue problems, we realize that this value is the square root of the largest eigenvalue of $A^\top A$, or equivalently

$$\|A\|_2 = \max\{\sigma_i\}.$$

In other words, we can read the two-norm of A directly from its eigenvalues.

Similarly, recall that the condition number of A is given by cond $A = \|A\|_2\|A^{-1}\|_2$. By our derivation of A^+, the singular values of A^{-1} must be the reciprocals of the singular values of A. Combining this with our simplification of $\|A\|_2$ yields:

$$\text{cond } A = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

This expression yields a strategy for evaluating the conditioning of A. Of course, computing $\sigma_{\min}$ requires solving systems $A\vec{x} = \vec{b}$, a process which in itself may suffer from poor conditioning of A; if this is an issue, conditioning can be bounded and approximated by using various approximations of the singular values of A.

6.2.4 The Procrustes Problem and Alignment

Many techniques in computer vision involve the alignment of three-dimensional shapes. For instance, suppose we have a three-dimensional scanner that collects two point clouds of the same rigid object from different views. A typical task might be to align these two point clouds into a single coordinate frame.

Since the object is rigid, we expect there to be some rotation matrix R and translation $\vec{t} \in \mathbb{R}^3$ such that that rotating the first point cloud by R and then translating by $\vec{t}$ aligns the two data sets. Our job is to estimate these two objects.

If the two scans overlap, the user or an automated system may mark n corresponding points that correspond between the two scans; we can store these in two matrices $X_1, X_2 \in \mathbb{R}^{3\times n}$. Then, for each column $\vec{x}_{1i}$ of X_1 and $\vec{x}_{2i}$ of X_2, we expect $R\vec{x}_{1i} + \vec{t} = \vec{x}_{2i}$. We can write an energy function measuring how much this relationship holds true:

$$E \equiv \sum_i \|R\vec{x}_{1i} + \vec{t} - \vec{x}_{2i}\|^2.$$

If we fix R and minimize with respect to $\vec{t}$, optimizing E obviously becomes a least-squares problem. Now, suppose we optimize for R with $\vec{t}$ fixed. This is the same as minimizing $\|RX_1 - X_2^t\|_{\text{Fro}}$, where the columns of X_2^t are those of X_2 translated by $\vec{t}$, subject to R being a 3×3 rotation matrix, that is, that $R^\top R = I_{3\times 3}$. This is known as the *orthogonal Procrustes problem.*

To solve this problem, we will introduce the *trace* of a square matrix as follows:

Definition 6.3 (Trace)**.** *The* trace *of $A \in \mathbb{R}^{n\times n}$ is the sum of its diagonal:*

$$tr(A) \equiv \sum_i a_{ii}.$$

It is straightforward to check that $\|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A)$. Thus, we can simplify E as follows:

$$\begin{aligned}\|RX_1 - X_2^t\|_{\text{Fro}}^2 &= \text{tr}((RX_1 - X_2^t)^\top (RX_1 - X_2^t)) \\ &= \text{tr}(X_1^\top X_1 - X_1^\top R^\top X_2^t - X_2^{t\top} RX_1 + X_2^{t\top} X_2) \\ &= \text{const.} - 2\text{tr}(X_2^{t\top} RX_1) \\ &\qquad \text{since } \text{tr}(A+B) = \text{tr}\, A + \text{tr}\, B \text{ and } \text{tr}(A^\top) = \text{tr}(A)\end{aligned}$$

Thus, we wish to maximize $\text{tr}(X_2^{t\top} RX_1)$ with $R^\top R = I_{3\times 3}$. In the exercises, you will prove that $\text{tr}(AB) = \text{tr}(BA)$. Thus our objective can simplify slightly to $\text{tr}(RC)$ with $C \equiv X_1 X_2^{t\top}$. Applying the SVD, if we decompose $C = U\Sigma V^\top$ then we can simplify even more:

$$\begin{aligned}\text{tr}(RC) &= \text{tr}(RU\Sigma V^\top) \text{ by definition} \\ &= \text{tr}((V^\top RU)\Sigma) \text{ since } \text{tr}(AB) = \text{tr}(BA) \\ &= \text{tr}(\tilde{R}\Sigma) \text{ if we define } \tilde{R} = V^\top RU, \text{ which is also orthogonal} \\ &= \sum_i \sigma_i \tilde{r}_{ii} \text{ since } \Sigma \text{ is diagonal}\end{aligned}$$

Since $\tilde{R}$ is orthogonal, its columns all have unit length. This implies that $\tilde{r}_{ii} \leq 1$, since otherwise the norm of column i would be too big. Since $\sigma_i \geq 0$ for all i, this argument shows that we can maximize $\text{tr}(RC)$ by taking $\tilde{R} = I_{3\times 3}$. Undoing our substitutions shows $R = V\tilde{R}U^\top = VU^\top$.

More generally, we have shown the following:

Theorem 6.2 (Orthogonal Procrustes). *The orthogonal matrix R minimizing $\|RX - Y\|^2$ is given by $VU^\top$, where SVD is applied to factor $XY^\top = U\Sigma V^\top$.*

Returning to the alignment problem, one typical strategy is an *alternating* approach:

1. Fix R and minimize E with respect to $\vec{t}$.

2. Fix the resulting $\vec{t}$ and minimize E with respect to R subject to $R^\top R = I_{3\times 3}$.

3. Return to step 1.

The energy E decreases with each step and thus converges to a local minimum. Since we never optimize $\vec{t}$ and R simultaneously, we cannot guarantee that the result is the smallest possible value of E, but in practice this method works well.

6.2.5 Principal Components Analysis (PCA)

Recall the setup from §5.1.1: We wish to find a low-dimensional approximation of a set of data points, which we can store in a matrix $X \in \mathbb{R}^{n\times k}$, for k observations in n dimensions. Previously, we showed that if we are allowed only a single dimension, the best possible direction is given by the dominant eigenvector of $XX^\top$.

Suppose instead we are allowed to project onto the span of d vectors with $d \leq \min\{k, n\}$ and wish to choose these vectors optimally. We could write them in an $n \times d$ matrix C; since we can apply Gram-Schmidt to any set of vectors, we can assume that the columns of C are orthonormal, showing $C^\top C = I_{d\times d}$. Since C has orthonormal columns, by the normal equations the projection of X onto the column space of C is given by $CC^\top X$.

In this setup, we wish to minimize $\|X - CC^\top X\|_{\text{Fro}}$ subject to $C^\top C = I_{d\times d}$. We can simplify our problem somewhat:

$$\begin{aligned}
\|X - CC^\top X\|_{\text{Fro}}^2 &= \text{tr}((X - CC^\top X)^\top (X - CC^\top X)) \text{ since } \|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A)\\
&= \text{tr}(X^\top X - 2X^\top CC^\top X + X^\top CC^\top CC^\top X)\\
&= \text{const.} - \text{tr}(X^\top CC^\top X) \text{ since } C^\top C = I_{d\times d}\\
&= -\|C^\top X\|_{\text{Fro}}^2 + \text{const.}
\end{aligned}$$

So, equivalently we can maximize $\|C^\top X\|_{\text{Fro}}^2$; for statisticians, this shows when the rows of X have mean zero that we wish to maximize the variance of the projection $C^\top X$.

Now, suppose we factor $X = U\Sigma V^\top$. Then, we wish to maximize $\|C^\top U\Sigma V^\top\|_{\text{Fro}} = \|\tilde{C}^\top \Sigma\|_{\text{Fro}} = \|\tilde{\Sigma}^\top C\|_{\text{Fro}}$ by orthogonality of V if we take $\tilde{C} = CU^\top$. If the elements of $\tilde{C}$ are $\tilde{c}_{ij}$, then expanding this norm yields

$$\|\Sigma^\top \tilde{C}\|_{\text{Fro}}^2 = \sum_i \sigma_i^2 \sum_j \tilde{c}_{ij}^2.$$

By orthogonality of the columns of $\tilde{C}$, we know $\sum_i \tilde{c}_{ij}^2 = 1$ for all j and, since $\tilde{C}$ may have fewer than n columns, $\sum_j \tilde{c}_{ij}^2 \leq 1$. Thus, the coefficient next to σ_i^2 is at most 1 in the sum above, so if we sort such that $\sigma_1 \geq \sigma_2 \geq \cdots$, then clearly the maximum is achieved by taking the columns of $\tilde{C}$ to be $\vec{e}_1, \ldots, \vec{e}_d$. Undoing our change of coordinates, we see that our choice of C should be the first d columns of U.

We have shown that the SVD of X can be used to solve such a *principal components analysis* (PCA) problem. In practice the rows of X usually are shifted to have mean zero before carrying out the SVD; as shown in Figure NUMBER, this centers the dataset about the origin, providing more meaningful PCA vectors $\vec{u}_i$.

6.3 Problems