8.2 How does this all work?[2]

Recall $\mathbf{A}$ is $m \times n$ so that $\mathbf{A}^T\mathbf{A}$ is $n \times n$ and $\mathbf{A}\mathbf{A}^T$ is $m \times m$. Both of these product matrices are square and symmetric.

So, by the result we saw earlier, $\mathbf{A}^T\mathbf{A}$ has n real eigenvalues and a set of n orthonormal eigenvectors (similarly for $\mathbf{A}\mathbf{A}^T$ which has m of them).

Let $\mathbf{v}_i$ be the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and λ_i be the corresponding eigenvectors and order them so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. (It can be shown that all eigenvales here are ≥ 0.)

Similarly, let $\mathbf{u}_i$ be the eigenvectors of $\mathbf{A}^T\mathbf{A}$ and μ_i be the corresponding eigenvectors and order them so that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \geq 0$.

It turns out that the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ are exactly the same as the non-zero eigenvalues of $\mathbf{A}^T\mathbf{A}$. Suppose there are r such non-zero eigenvalues, so that $\lambda_{r+1} = \ldots = \lambda_n = 0$ and $\mu_{r+1} = \ldots = \mu_m = 0$.

r is called the *rank* of $\mathbf{A}$ (and of $\mathbf{A}^T$). Clearly, $r \leq m$ and $r \leq n$.

Now, for $k = 1, \ldots, r$, it can be shown that we have

$$\mathbf{A}\mathbf{v}_k = \sigma_k \mathbf{u}_k \text{ and } \mathbf{A}^T\mathbf{u}_k = \sigma_k \mathbf{v}_k$$

where $\sigma_k = \sqrt{\lambda}_k = \sqrt{\mu}_k$. And, also that, for $k > r$,

$$\mathbf{A}\mathbf{v}_k = 0 \text{ and } \mathbf{A}^T\mathbf{u}_k = 0.$$

The equations $\mathbf{A}\mathbf{v}_k = \sigma_k\mathbf{u}_k$ for $k \leq r$ together with $\mathbf{A}\mathbf{v}_k = 0$ for $k > r$ tell us how $\mathbf{A}$ acts on the orthonormal set of vectors $\{\mathbf{v}_k\}$. Since this set is a basis for $\mathbb{R}^n$, the equations give a complete description of the action of $\mathbf{A}$, so that we can write

$$\mathbf{A} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}. \tag{5}$$

A similar argument shows that

$$\mathbf{A}^{\mathrm{T}} = \sum_{k=1}^{r} \sigma_k \mathbf{v}_k \mathbf{u}_k^{\mathrm{T}}.$$

The orthonormal vectors $\{\mathbf{v}_k\}$ are known as the *right singular vectors*, the vectors $\{\mathbf{u}_k\}$ are known as the *left singular vectors*, and the scalars $\{\sigma_k\}$ are called the *singular values* of the matrix $\mathbf{A}$.

The singular value decomposition allows us to understand the action of $\mathbf{A}$ on a vector $\mathbf{x}$ as

$$\mathbf{A}\mathbf{x} = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k (\mathbf{v}_k^{\mathrm{T}} \mathbf{x}).$$

which can be interpreted as having three stages:

[2]This section is taken, and condensed, from notes written by Sze Tan for Physics 707: Inverse Problems.

1. It resolves the input vector along each of the right singular vectors $\mathbf{v}_k$, the component of the input vector along the kth singular vector being given by $\mathbf{v}_k^T\mathbf{x}$,
2. The amount along the kth direction is multiplied by the singular value σ_k,
3. The product tells us how much of the kth left singular vector $\mathbf{u}_k$ is present in the product $\mathbf{Ax}$.

This is illustrated in Figure 4.

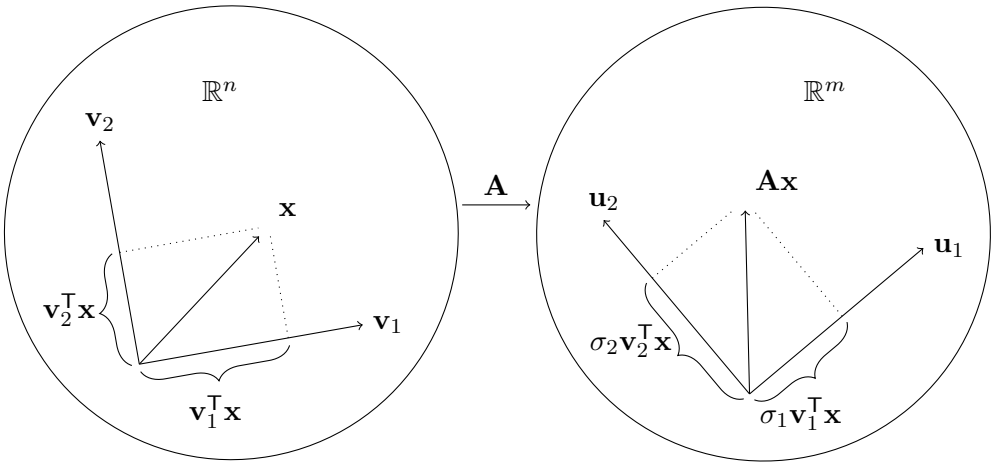

Figure 4: Effect of a rectangular matrix $\mathbf{A}$ of size $m \times n$ on a vector $\mathbf{x}$. Only two of the orthogonal eigenvectors are shown.

8.3 Structure of SVD

In the overdetermined case, in which $m > n$, so that we have more equations than unknowns, we have the following structure:

$$\mathbf{A} = \mathbf{U}\ \mathbf{D}\ \mathbf{V}^\top$$

In the underdetermined case, in which $m < n$, so that we have fewer equations than unknowns, we have the following structure:

$$\mathbf{A} = \mathbf{U}\ \mathbf{D}\ \mathbf{V}^\top$$

Note that, in the overdetermined case, we truncate $\mathbf{U}$ and $\mathbf{D}$ since there are at most $r \leq n < m$ non-zero singular values of $\mathbf{A}$ we can omit the $\mathbf{u}_i$ that contribute nothing to matrix product.

The matrix $\mathbf{V}$ is orthonormal, so $\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$. $\mathbf{U}$ is orthonormal when $m \geq n$, but if $m < n$, the singular values $\sigma_j = 0$ for $j = m+1, \ldots, n$ and the corresponding

columns of $\mathbf{U}$ are also 0 so that $\mathbf{U}\mathbf{U}^T = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$ where only the first m elements of the diagonal are 1 and the elements from $m+1$ to n are zero.

Note that the SVD of matrix $\mathbf{A}$ is only unique up to permutations of the columns/rows. For this reason, we insist that the singular values and corresponding singular vectors are arranged so that the singular values are in descending order $\sigma_1 \geq \sigma_2 \geq \ldots$. Even then, some of the σ_i's may have the same value so columns of $\mathbf{U}$ and $\mathbf{V}$ could be permuted. Aside from these possible permuations, the representation is unique. Be aware when calculating the SVD with various software that the you may need to enforce this canonical representation.

8.4 Condition number of a matrix

The concept of a condition number was introduced in Section 2. This concept can be applied to a matrices and is useful, for example, when considering solutions to the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$. Solutions to this equation will change greatly with small changes in $\mathbf{b}$ when $\mathbf{A}$ has a large condition number, while the small changes in $\mathbf{b}$ will lead to only small changes in the solution when the matrix has a small condition number. We can define the condition number of a matrix as the maximum of the ratio of the relative error in $\mathbf{x}$ divided by the relative error in $\mathbf{b}$, where the maximum is taken over all possible $\mathbf{x}$ and $\mathbf{b}$.

To give a full description of how to derive the condition number of $\mathbf{A}$, we would have to introduce matrix norms which we do not have time to do here. Instead, we simply present the result here that the condition number of $\mathbf{A}$ can be defined as the ratio of the largest to the smallest non-zero singular values:

$$cond(\mathbf{A}) = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

If the smallest singular value of $\mathbf{A}$ is 0, $\mathbf{A}$ is singular (has no inverse) but the condition number of $\mathbf{A}$ is still defined.

The condition number of $\mathbf{A}$ is considered to be large, and the matrix is *ill-conditioned*, if roughly $\log(cond(\mathbf{A})) \geq k$ where k is the number of digits of precision in the matrix entries.

Example: Find the condition number of the matrix $\mathbf{A} = \begin{bmatrix} 2 & -3 \\ 1 & -1 \end{bmatrix}$

Solution: Using a matrix algebra package, find the singular values of $\mathbf{A}$ to be 3.864 and 0.259, so $cond(\mathbf{A}) = \frac{3.864}{0.259} \approx 14.9$. □

Example: The singular values of the matrix $\mathbf{A} = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}$ are approximately 1.58 and 6.33×10^{-9} so the condition number is about 2.5×10^8. This very large condition number means that $\mathbf{A}$ is an ill-conditioned matrix.

Ill-conditioning means that standard approaches to solving linear systems can be very unstable. For example, consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{b} = \begin{bmatrix} 0.8642 \\ 0.1140 \end{bmatrix}$. This

has the exact solution $\mathbf{x} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$.

But standard matrix software (`linsolve` in Matlab), gives the solution as $\begin{bmatrix} 2.59 \\ -3.89 \end{bmatrix} \times 10^6$ which is radically wrong!

It also means that if some number in the system is measured slightly differently, the results we get can change enormously. For example, if the measurement vector $\mathbf{b}$ is just slightly different, say $\mathbf{b} = \begin{bmatrix} 0.86419999 \\ 0.11400001 \end{bmatrix}$, then the exact solution is now close to $\mathbf{x} = \begin{bmatrix} 0.9911 \\ -0.4870 \end{bmatrix}$ which represents an enormous change in the solution relative to the small change in the original system. □

8.5 Applications of SVD

We'll see in a later section on pseudo-inverses how the SVD can be used to solve the linear system $\mathbf{Ax} = \mathbf{b}$.

8.5.1 Image compression

See slides and assignment 1.

8.5.2 Gene expression

Abstract from relevant paper: We describe the use of singular value decomposition in transforming genome-wide expression data from genes × arrays space to reduced diagonalized eigengenes × eigenarrays space, where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). Normalizing the data by filtering out the eigengenes (and eigenarrays) that are inferred to represent noise or experimental artifacts enables meaningful comparison of the expression of different genes across different arrays in different experiments. Sorting the data according to the eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. After normalization and sorting, the significant eigengenes and eigenarrays can be associated with observed genome-wide effects of regulators, or with measured samples, in which these regulators are overactive or underactive, respectively.