Lecture 13: Singular value decomposition

Recap

Last time we reviewed some basics of linear algebra and eigenvalue problems, and we finished up with a classic eigenvalue equation for a rank 2 matrix with 2 modes. We finished up with an eigenvalue decomposition of the form:

$$\mathbf{A}^{-1} = \mathbf{P}\mathbf{D}^{-1}\mathbf{P}^T \tag{1}$$

where $\mathbf{A}$ is a square matrix, $\mathbf{P}$ is an orthonormal matrix containing a set of basis vectors that span the space defined by $\mathbf{A}$, and $\mathbf{D}$ is a diagonal matrix with the eigenvalues on the diagonal.

The *condition number* of the matrix $\mathbf{A}$ is the ratio of the largest to the smallest eigenvalue and is an indication of the stability of the inversion to numerical error.

Representing matrices that are not square

Standard eigenvalue calculations make sense for square matrices, but what happens for a matrix $\mathbf{G}$ that is not square? Let a linear transformation be defined by the $N \times M$ matrix $\mathbf{G}$, so that $\mathbf{Gm}$ is a transformation of $\mathbf{m}$ into $\mathbf{d}$. Define $N \times N$ and $M \times M$ orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ by some basis vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ as

$$\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots \mathbf{u}_N\end{bmatrix} \tag{2}$$
$$\mathbf{V} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots \mathbf{v}_N\end{bmatrix}. \tag{3}$$

The vector $\mathbf{m}$ is transformed into $\tilde{\mathbf{m}}$ through a matrix rotation operation.

$$\tilde{\mathbf{m}} = \mathbf{V}^T\mathbf{m}. \tag{4}$$

That is, $\tilde{m}$ is the coordinates of $\mathbf{m}$ with respect to the basis vectors $\mathbf{v}_i$. Because $\mathbf{V}$ is orthogonal, the inverse transform is

$$\mathbf{m} = \mathbf{V}\tilde{\mathbf{m}}. \tag{5}$$

Similarly for vector $\mathbf{d}$

$$\tilde{\mathbf{d}} = \mathbf{U}^T\mathbf{d} \tag{6}$$
$$\mathbf{d} = \mathbf{U}\tilde{\mathbf{d}}. \tag{7}$$

Consider the misfit vector

$$\mathbf{e} = \mathbf{Gm} - \mathbf{d} \tag{8}$$

Using (5) and (7)

$$\mathbf{e} = \mathbf{GV}\tilde{\mathbf{m}} - \mathbf{U}\tilde{\mathbf{d}} \tag{9}$$

Premultiplying by $\mathbf{U}^T$:

$$\mathbf{U}^T\mathbf{e} = \mathbf{U}^T\mathbf{GV}\tilde{\mathbf{m}} - \tilde{\mathbf{d}}. \tag{10}$$

So that the transformed misfit vector is

$$\tilde{\mathbf{e}} = \mathbf{U}^T\mathbf{GV}\tilde{\mathbf{m}} - \tilde{\mathbf{d}}. \tag{11}$$

The linear transformation $\mathbf{G}$ is represented in the new basis by

$$\tilde{\mathbf{G}} = \mathbf{U}^T\mathbf{GV}. \tag{12}$$

Everything we've done so far has been hypothetical. We haven't relied on any special knowledge of $\mathbf{G}$, or any unusual requirements for $\mathbf{U}$ or $\mathbf{V}$.

Our goal is to find two orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ such that $\tilde{\mathbf{G}}$ is as simple as possible. Let's assume that $\tilde{\mathbf{G}} = \mathbf{S}$ where

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{13}$$

where $\mathbf{S}_K$ is a $K \times K$ diagonal matrix

$$\mathbf{S}_K = \begin{bmatrix} s_1 & & & \mathbf{0} \\ & s_2 & & \\ & & \ddots & \\ \mathbf{0} & & & {}_K \end{bmatrix}. \tag{14}$$

Such a representation for $\tilde{\mathbf{G}}$ would certainly be very simple.

Now we show that such a representation is always possible. From (12)

$$\mathbf{G} = \mathbf{U}\mathbf{S}\mathbf{V}^T. \tag{15}$$

So the symmetric matrix $\mathbf{G}^T\mathbf{G}$ is

$$\mathbf{G}^T\mathbf{G} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}(\mathbf{S}^T\mathbf{S})\mathbf{V}^T, \tag{16}$$

where $\mathbf{S}^T\mathbf{S}$ is diagonal. Since $\mathbf{V}$ is orthogonal, we know that the diagonal elements of $\mathbf{S}^T\mathbf{S}$ are the eigenvalues of $\mathbf{G}^T\mathbf{G}$. Because $\mathbf{G}^T\mathbf{G}$ is symmetric, we know that such an eigenvalue decomposition is always possible. Similarly

$$\mathbf{G}\mathbf{G}^T = \mathbf{U}(\mathbf{S}\mathbf{S}^T)\mathbf{U}^T, \tag{17}$$

where $\mathbf{S}\mathbf{S}^T$ is diagonal, and the diagonal elements are the eigenvalues of $\mathbf{G}\mathbf{G}^T$. The elements along the diagonal of $\mathbf{S}$ are called the singular values of $\mathbf{G}$. It may be surprising (but true) that the singular value decomposition (15) is possible for any matrix $\mathbf{G}$. The number K of non-zero singular values is the rank of $\mathbf{G}$.

We are now in the position to prove a few things about the general inverse problem. First, we define the transformed vectors as being separated into two parts

$$\begin{aligned} \mathbf{U}^T\mathbf{d} &= \tilde{\mathbf{d}} = \begin{bmatrix} \tilde{\mathbf{d}}_K \\ \tilde{\mathbf{d}}_0 \end{bmatrix} && (18) \\ \mathbf{V}^T\mathbf{m} &= \tilde{\mathbf{m}} = \begin{bmatrix} \tilde{\mathbf{m}}_K \\ \tilde{\mathbf{m}}_0 \end{bmatrix} && (19) \\ & && (20) \end{aligned}$$

consisting of the first K components, and the remaining components. All solutions to the problem of minimizing $||\mathbf{G}\mathbf{m} - \mathbf{d}||_2$ are of the form

$$\mathbf{m} = \mathbf{V}\tilde{\mathbf{m}} = \mathbf{V}\begin{bmatrix} \tilde{\mathbf{m}}_K \\ \tilde{\mathbf{m}}_0 \end{bmatrix} \tag{21}$$

where $\tilde{\mathbf{m}}_0$ is arbitrary. Prove this by considering (dropping the 2 subscript from here on)

$$\begin{aligned} ||\mathbf{G}\mathbf{m} - \mathbf{d}||_2^2 &= ||\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{m} - \mathbf{d}||^2 && (22) \\ &= ||\mathbf{S}\mathbf{V}^T\mathbf{m} - \mathbf{U}^T\mathbf{d}||^2 && (23) \\ &= ||\mathbf{S}\tilde{\mathbf{m}} - \tilde{\mathbf{d}}||^2 && (24) \\ &= ||\mathbf{S}_K\tilde{\mathbf{m}}_K - \tilde{\mathbf{d}}_K||^2 + ||\tilde{\mathbf{d}}_0||^2. && (25) \end{aligned}$$

The minimum misfit is for

$$\mathbf{S}_K \tilde{\mathbf{m}}_K = \tilde{\mathbf{d}}_K, \tag{26}$$

and the misfit is then $||\tilde{\mathbf{d}}_0||^2$. Apparently $\tilde{\mathbf{m}}_0$ has no effect on the misfit. Thus it is in the null space. For the full rank underdetermined problem, there is no $\tilde{\mathbf{d}}_0$, and therefore no misfit. For the full rank overdetermined problem, there is no $\tilde{\mathbf{m}}_0$ and therefore no null space. The solution with minimum model size as measured by $||\mathbf{m}||$ is clearly the one with $\tilde{\mathbf{m}}_0 = \mathbf{0}$.

Generalized inverse

We can define the generalized inverse (or pseudoinverse, or Moore-Penrose inverse). The unique minimum length solution to the inverse problem of minimizing $||\mathbf{Gm} - \mathbf{d}||$ is

$$\mathbf{m} = \mathbf{G}^+\mathbf{d}, \tag{27}$$

where

$$\mathbf{G}^+ = \mathbf{V}\mathbf{S}^+\mathbf{U}^T \tag{28}$$

$$\mathbf{S}^+ = \begin{bmatrix} \mathbf{S}_K^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{29}$$

Note that $\mathbf{S}_K$ is easy to invert as it is a diagonal matrix. This is a suitable inverse for you if your idea of the best model is the one with minimum length as measured by the L_2 norm of $\mathbf{m}$.

Practical use of the singular value decomposition

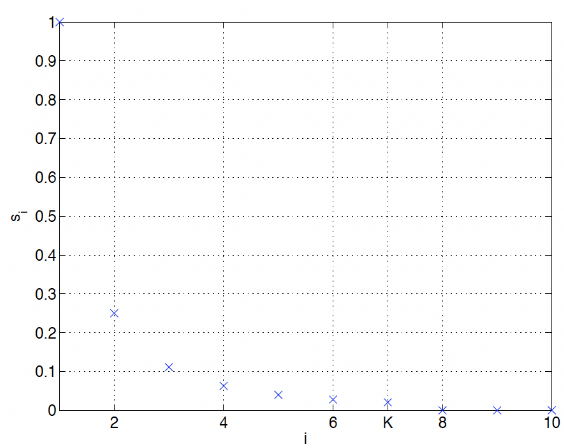

Figure 1: Singular values of a matrix of rank K.

Singular value decompositions have some powerful uses for data analysis, both for carrying out matrix inversion and also for assessing variability in general. Here are some things to consider.

1. *Rank.* The singular value decomposition is used primarily because it tells us immediately about the rank of the problem. Typically, one plots the singular values against their index numbers (Figure 1). This is often called the "spectrum" of the data kernel. The first K singular values are nonzero, but some of the nonzero singular values are very small. In practice, some of these small singular values are set to zero, increasing misfit, but reducing model size. This sounds similar to the procedure we used in simultaneously minimizing misfit and model size, because it is essentially the same thing.

2. *Minimizing the model size.* Consider the generalized inverse (27), for which the model size is

$$||\tilde{\mathbf{m}}_K||^2 = ||\mathbf{S}_K^{-1}\tilde{d}_K||^2 = \sum_{i=1}^{K} \left(s_i^{-1}\tilde{d}_i \right)^{-2} \tag{30}$$

3. *Characterizing misfit.* The misfit is contained within the null space of the matrix and is

$$||\tilde{\mathbf{d}}_0||^2 = \sum_{i=K+1}^{N} d_i^2 \tag{31}$$

   As small singular values are set to zero, the model size will decrease, while the misfit increases.

4. *Assessing the solution covariance.* Recall the model covariance matrix, which for the generalized inverse is

$$\langle \mathbf{m}'\mathbf{m}'^T \rangle = \mathbf{G}^+ \langle \mathbf{d}'\mathbf{d}'^T \rangle \mathbf{G}^{+T}. \tag{32}$$

   Substituting the identity matrix for the data covariance matrix yields the **unit covariance matrix**:

$$\langle \mathbf{m}'\mathbf{m}'^T \rangle_u = \mathbf{G}^+ \mathbf{G}^{+T}. \tag{33}$$

   Using (28-29) the unit covariance matrix can be written

$$\begin{aligned}
\langle \mathbf{m}'\mathbf{m}'^T \rangle_u &= \mathbf{V}\mathbf{S}^+\mathbf{U}^T\mathbf{U}\mathbf{S}^{+T}\mathbf{V}^T & (34)\\
&= \mathbf{V} \begin{bmatrix} \mathbf{S}_K^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{S}_K^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T & (35)\\
&= \mathbf{V} \begin{bmatrix} \mathbf{S}_K^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T & (36)\\
&= \mathbf{V}_K \mathbf{S}_K^{-2} \mathbf{V}_K^T, & (37)
\end{aligned}$$

   where $\mathbf{V}_K$ is the matrix made of the first K columns of $\mathbf{V}$. The model variance is dominated by the smallest singular values. The fact that the small singular values dominate the covariance of $\mathbf{m}$ tells us that small, erroneous singular values could corrupt our estimates of model paramters. This is one more reason to zero out the small singular values, and move those components of the solution to the null space.

Using singular values to assess dominate modes of variability

The singular value decomposition provides us with a framework for assessing the dominant patterns of variability in any data set, provided that we can represent the data in a matrix form—typically with time in one matrix dimension and space in the other dimension. In class, we looked at examples of checkerboard patterns to gain insight about compact representation of data with a singular value decomposition.

Example 1: What will the singular value decomposition look like for a checkerboard pattern, as shown in Figure 2?

A checkerboard has clear alternating patterns. Every row is easily represented by as either +1 or -1 times the other rows. Such as system has one non-zero singular value, one meaningful vector $\mathbf{u}_1$, and one meaningful vector $\mathbf{v}_1$.

Example 2: What happens when the pattern is more complicated, as in Figure 3?

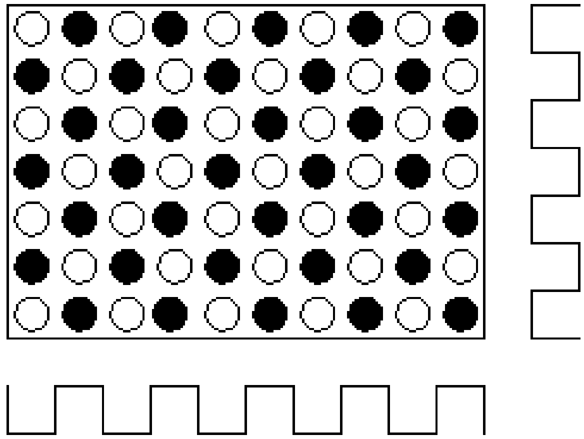

Figure 2: Checkerboard pattern. You could imagine this as a data set consising of alternating values of +1 and -1.

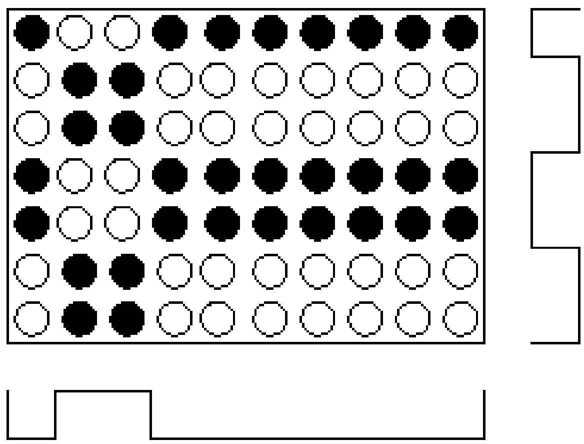

Figure 3: Structured pattern of variability.

In this case, although the patterns are not predictability repeated, each row can be represented as either +1 or -1 times the other rows, and each column is +1 or -1 times the other columns. Again the system has one non-zero singular value, one meaningful vector $\mathbf{u}_1$, and one meaningful vector $\mathbf{v}_1$.

Example 3: What happens for a propagating signal, as in Figure 4?

In this case, the rows and columns are not all identical. There are effectively two types of rows and two types of columns, so 2 non-zero singular values. This is our reminder that propagating patterns necessarily are represented by a singular value decomposition as paired orthogonal patterns (a sine mode and a cosine mode).

Example 4: What happens for a a highly unstructured signal, as in Figure 5?

In this case, rows and columns are not trivially represented as linear combinations of each other, and the singular value decomposition will have a high rank (and little compactness in representing observed variability).

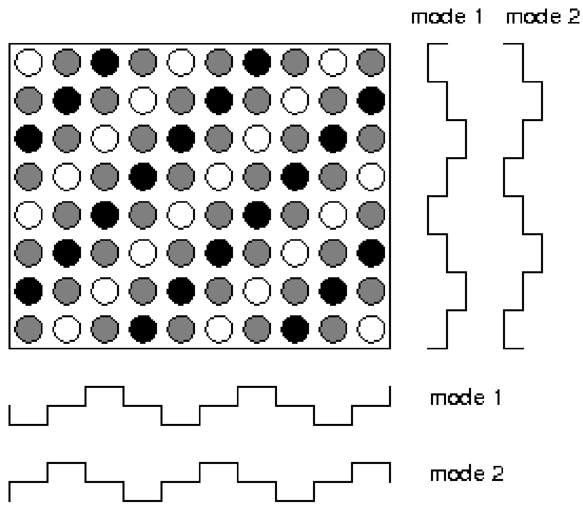

Figure 4: Propagating signal. You can imagine this to consist of values of +1, 0, and -1, depending on color.

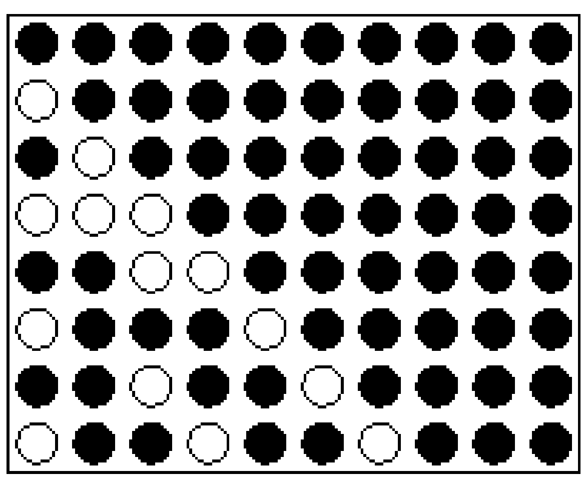

Figure 5: Unstructured signal.