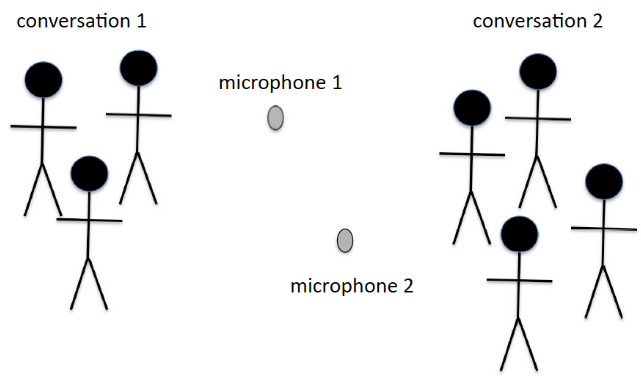
Lecture 15
Independent Component Analysis and Image Separation

Jason J. Bramburger

Let's start with a simple way to intuitively understand the concept of independent components. Imagine you're at a gathering and in one room there are two concurrent conversations taking place simultaneously. Furthermore, let's suppose there are two microphones set up in the room at different spatial locations so that they can pick up audio from both conversations. We denote the signals from each conversation as $s_1(t)$ and $s_2(t)$ and ask the following question: How can we separate the signals that have been mixed at each of the microphone locations?



From a mathematical perspective, this problem can be formulated with the following mixing (linear combination) equation

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$$

where $x_1(t)$ and $x_2(t)$ are the mixed, recorded signals at each microphone. The coefficients a_{ij} are the mixing parameters that are determined by a variety of factors including the placement of the microphones in the room, the distance to the conversations, and the overall room acoustics. Our goal then becomes recovering the **independent components** $s_1(t)$ and $s_2(t)$ assuming we are only given $x_1(t)$ and $x_2(t)$.

<u>**Mathematical Framework**</u>

The general setting framework for **Independent Component Analysis** (ICA) is established in a similar manner to the above example. Suppose we are given N distinct linear combinations of N signals or data sets. Then, we would like to determine the original N signals or data sets. Mathematically, we are given $x_j(t)$ such that

$$x_j(t) = a_{j1}s_1(t) + a_{j2}s_2(t) + \cdots + a_{jN}s_N(t), \quad 1 \le j \le N.$$

We can write this compactly in matrix form as

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

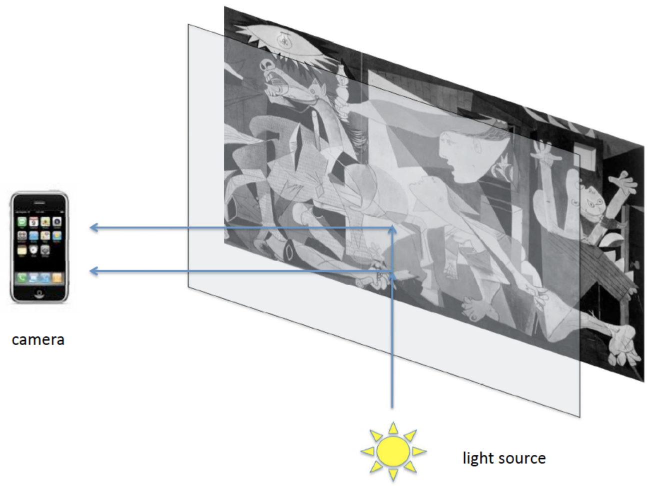
where $\mathbf{x}$ are the mixed signal measurements, and $\mathbf{s}$ are the original signals we are trying to determine. On first glance, you probably think this problem is easy since we could just invert $\mathbf{A}$ to get

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}.$$

Doing this though makes the following assumption: we know the matrix $\mathbf{A}$. This problem is really only interesting if we don't know the coefficients a_{ij} that make up the matrix $\mathbf{A}$.

Image Separation and SVD

Let's turn our attention to the problem of image separation. A prototypical example is that you are photographing an image behind glass, as illustrated below:



The glass encasing the image produces a reflection of the backlit source along with the image itself. Your eyes already can separate out the independent components of this image since you can see through much of the reflection from the glass on the original image. Unfortunately, a computer cannot do this and so we are interested in training a computer to do this image processing automatically. This constitutes what is called *artificial vision* with the goal being to mimic natural human behaviours and abilities.

Since there are two images (the actual image and the reflection) that need to be teased apart, we are going to two two images to work with. To separate the images, a second photo is taken as illustrated above but now with a linear polarizer placed in front of the camera lens. Now, we have two images $\mathbf{S}_1$ and $\mathbf{S}_2$ which are unknown, along with the mixing coefficients $(a_{11}, a_{12}, a_{21}, a_{22})$ which make up the matrix $\mathbf{A}$. The only thing we have are the two pictures we took: $\mathbf{X}_1$ and $\mathbf{X}_2$. We can't solve this system in general without some assumptions being made first. They are as follows:

1. The two images $\mathbf{S}_1$ and $\mathbf{S}_2$ are statistically independent. Mathematically, this means that

$$P(\mathbf{S}_1, \mathbf{S}_2) = P(\mathbf{S}_1)P(\mathbf{S}_2),$$

   which states that the joint probability density is separable. No specific form is placed upon the probability densities except that they cannot be Gaussian[1].

2. The matrix $\mathbf{A}$ is full rank. This amounts to assuming that the two measurements of the image have indeed been filtered with different polarization states.

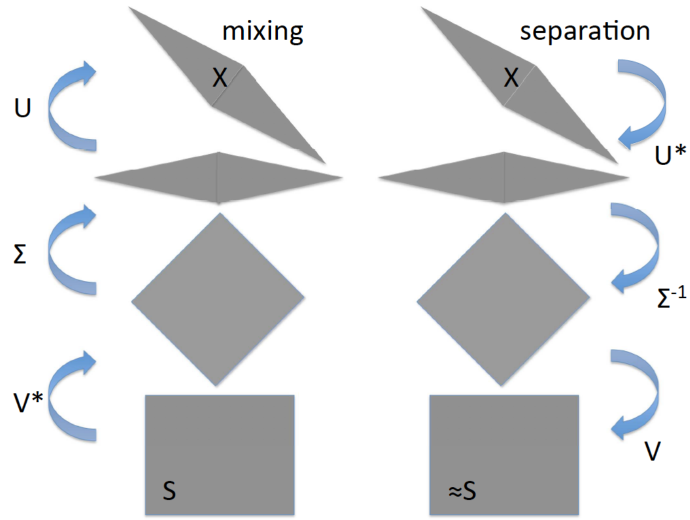
Let's just think about what happens when we apply the SVD to the matrix $\mathbf{A}$. We get

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*,$$

where we recall that $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices that simply lead to rotation and Σ scales an image as prescribed by the singular values. Below is a graphical illustration of this process for distributions that are uniform, i.e. they form a square. The mixing matrix $\mathbf{A}$ can be thought of as first rotating the square via the matrix $\mathbf{V}^*$, stretching it into a parallelogram with the diagonal matrix Σ, then rotating the parallelogram via the matrix $\mathbf{U}$. This is the process by which we arrive at the mixed image $\mathbf{X}$.

Big Picture: The estimation of the independent images reduces to finding how to transform the rotated parallelogram back into a square. Mathematically, we transform the mixed image back into a separable product of

[1] In practice, this is a reasonable constraint since natural images are rarely Gaussian.



one-dimensional probability distributions.

Above we described the method by which we move from the two independent images to the mixed images. Now, we will describe how to undo this process mathematically, following the right-hand-side of the above figure.

Step 1: Rotation of the Parallelogram

The first thing we want is to undo the rotation of the unitary matrix $\mathbf{U}$ by applying $\mathbf{U}^*$ to the mixed images. Our objective is to align the long and short axes of the parallelogram with the primary axis, as depicted in the top right of the above figure. The angle of the parallelogram relative to the primary axes will be denoted θ, and the long and short axes correspond to the axes of the maximal and minimal variance, respectively. From the data itself, we can obtain these directions. Assuming our measurements have mean zero, the variance at an arbitrary angle of orientation is given by

$$\mathrm{Var}(\theta) = \sum_{j=1}^{N} \left(\begin{bmatrix} x_1(j) & x_2(j) \end{bmatrix} \cdot \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \right)^2$$

Above we have a function of θ, so we can find the maximum using basic calculus. We can then find the minimum by assuming that the corresponding value of θ will give an orthogonal direction to the maximum value. Hence, if θ_0 is the maximum, then $\theta_0 - \frac{\pi}{2}$ will be the minimum.

To begin, we will expand out the dot product and exponents in the variance function to arrive at

$$\mathrm{Var}(\theta) = \sum_{j=1}^{N} x_1^2(j)\cos^2(\theta) + 2x_1(j)x_2(j)\cos(\theta)\sin(\theta) + x_2^2(j)\sin^2(\theta).$$

Differentiating with respect to θ gives

$$\frac{\mathrm{dVar}}{\mathrm{d}\theta} = 2\sum_{j=1}^{N}[x_2^2(j) - x_1^2(j)]\cos(\theta)\sin(\theta) + x_1(j)x_2(j)[\cos^2(\theta) - \sin^2(\theta)],$$

for which we can use the double-angle trigonometric identities to simply this expression to

$$\frac{\mathrm{dVar}}{\mathrm{d}\theta} = 2\sum_{j=1}^{N}[x_2^2(j) - x_1^2(j)]\sin(2\theta) + x_1(j)x_2(j)\cos(2\theta).$$

Setting $\mathrm{dVar}/\mathrm{d}\theta = 0$ and rearranging then gives us

$$\underbrace{\frac{\sin(2\theta)}{\cos(2\theta)}}_{=\tan(2\theta)} = \frac{-2\sum_{j=1}^{N} x_1(j)x_2(j)}{\sum_{j=1}^{N}[x_2^2(j) - x_1^2(j)]}.$$

Inverting the tangent function gives

$$\theta_0 = \frac{1}{2}\tan^{-1}\left(\frac{-2\sum_{j=1}^{N} x_1(j)x_2(j)}{\sum_{j=1}^{N}[x_2^2(j) - x_1^2(j)]}\right).$$

Hence, the matrix $\mathbf{U}^*$ is given by

$$\mathbf{U}^* = \begin{bmatrix} \cos(\theta_0) & \sin(\theta_0) \\ -\sin(\theta_0) & \cos(\theta_0) \end{bmatrix}$$

and is computed directly from the data $x_1(t)$ and $x_2(t)$ to produce θ_0.

Step 2: Scaling of the Parallelogram

Next we want to undo the principal component scaling due to the singular values of in the SVD decomposition. Recall that above we assumed that along the direction θ_0 the maximal variance is achieved, while along $\theta_0 - \pi/2$ the minimal variance is achieved. Hence, the singular values are computed as

$$\sigma_1 = \sum_{j=1}^{N}\left(\begin{bmatrix} x_1(j) & x_2(j) \end{bmatrix} \cdot \begin{bmatrix} \cos(\theta_0) \\ \sin(\theta_0) \end{bmatrix}\right)^2 = \mathrm{Var}(\theta_0)$$

$$\sigma_2 = \sum_{j=1}^{N}\left(\begin{bmatrix} x_1(j) & x_2(j) \end{bmatrix} \cdot \begin{bmatrix} \cos(\theta_0 - \pi/2) \\ \sin(\theta_0 - \pi/2) \end{bmatrix}\right)^2 = \mathrm{Var}(\theta_0 - \pi/2).$$

The square roots of these values make up the diagonal components of Σ. So, to undo the scaling, we apply the inverse of Σ, given by

$$\Sigma^{-1} = \begin{bmatrix} 1/\sqrt{\sigma_1} & 0 \\ 0 & 1/\sqrt{\sigma_2} \end{bmatrix}.$$

Step 3: Rotation to Produce a Separable Probability Distribution

The only thing we have left to do now is to rotate in such a way to separate the probability distribution. This separation process typically relies on the higher moments of the probability distribution. Recall that we have assumed the mean to be zero, and we have already calculated the variances. We have no reason to believe that the probability distribution is asymmetric, meaning that the higher order odd moments are negligible. Hence, the next nonzero moment that we consider will be kurtosis (the fourth moment), since we have assumed that skewness (the third moment) is zero. Our goal will be to minimize kurtosis, and by doing so we will determine the appropriate rotation angle. Mathematically, minimizing the kurtosis will be another step in trying to approximate the probability distribution of the images as separable functions so that $P(\mathbf{S}_1, \mathbf{S}_2) \approx P(\mathbf{S}_1)P(\mathbf{S}_2)$.

Let us begin by denoting $\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1(t) & x_2(t) \end{bmatrix}$ to be the transformed original images using the previous steps. That is,

$$\bar{\mathbf{x}} = \Sigma^{-1}\mathbf{U}^*\mathbf{x}.$$

Then, the (normalized) kurtosis of the probability distribution is given by

$$\bar{K}(\phi) = \sum_{j=1}^{N} \frac{1}{\bar{x}_1^2(j) + \bar{x}_2^2(j)}\left(\begin{bmatrix} \bar{x}_1(j) & \bar{x}_2(j) \end{bmatrix} \cdot \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}\right)^4,$$

where ϕ is the angle of rotation associated with the unitary matrix $\mathbf{V}$. Since we want to minimize $\bar{K}$ with respect to ϕ, we can proceed in a similar manner to Step 1 above and fist expand out $\bar{K}$ and then take the derivative with respect to ϕ. I am going to skip some steps (they can be found in the textbook if you are curious) and jump to the following cleaned up first derivative of $\bar{K}$:

$$\frac{\mathrm{d}\bar{K}}{\mathrm{d}\phi} = \sum_{j=1}^{N}\Bigg([\bar{x}_1^2(j) - \bar{x}_2^2(j)]\sin(2\phi) + 2\bar{x}_1(j)\bar{x}_2(j)\cos(2\phi)$$
$$+ \frac{1}{\bar{x}_1^2(j) + \bar{x}_2^2(j)}[2\bar{x}_1^3(j)\bar{x}_2(j) - 2\bar{x}_1(j)\bar{x}_2^3(j)]\cos(4\phi) + [3\bar{x}_1^2(j)\bar{x}_2^2(j) - (1/2)\bar{x}_1^4(j) - (1/2)\bar{x}_2^4(j)]\sin(4\phi)\Bigg).$$

Notice that the $\cos(2\phi)$ and $\sin(2\phi)$ terms together form the derivative of the variance of $\bar{\mathbf{x}}$, as a function of ϕ. We would like to minimize both variance and kurtosis, and so by setting the above 2ϕ terms to zero we are doing both. Hence,

$$\frac{\mathrm{d}\bar{K}}{\mathrm{d}\phi} = \sum_{j=1}^{N}\left(\frac{1}{\bar{x}_1^2(j)+\bar{x}_2^2(j)}[2\bar{x}_1^3(j)\bar{x}_2(j) - 2\bar{x}_1(j)\bar{x}_2^3(j)]\cos(4\phi) + [3\bar{x}_1^2(j)\bar{x}_2^2(j) - (1/2)\bar{x}_1^4(j) - (1/2)\bar{x}_2^4(j)]\sin(4\phi)\right),$$

and setting $\mathrm{d}\bar{K}/\mathrm{d}\phi = 0$ allows us to isolate for ϕ so that

$$\phi_0 = \frac{1}{4}\tan^{-1}\left(\frac{-\sum_{j=1}^{N}[2\bar{x}_1^3(j)\bar{x}_2(j) - 2\bar{x}_1(j)\bar{x}_2^3(j)]/[\bar{x}_1^2(j)+\bar{x}_2^2(j)]}{\sum_{j=1}^{N}[3\bar{x}_1^2(j)\bar{x}_2^2(j) - (1/2)\bar{x}_1^4(j) - (1/2)\bar{x}_2^4(j)]/[\bar{x}_1^2(j)+\bar{x}_2^2(j)]}\right).$$

Therefore, the rotation back to the approximately statistically independent square is then given by

$$\mathbf{V} = \begin{bmatrix}\cos(\phi_0) & \sin(\phi_0)\\ -\sin(\phi_0) & \cos(\phi_0)\end{bmatrix}.$$

At this point, it is unknown whether the angle ϕ_0 is a minimum or a maximum of the kurtosis. This needs to be checked. Sometimes this is a maximum, especially in cases when one of the image histograms has long tails relative to the other.

Putting it All Together

We have seen that we can (approximately) reconstruct the image through the following linear algebra:

$$\begin{aligned}\mathbf{s} &= \mathbf{A}^{-1}\mathbf{x} = \mathbf{V}\Sigma^{-1}\mathbf{U}^*\mathbf{x}\\ &= \begin{bmatrix}\cos(\phi_0) & \sin(\phi_0)\\ -\sin(\phi_0) & \cos(\phi_0)\end{bmatrix}\begin{bmatrix}1/\sqrt{\sigma_1} & 0\\ 0 & 1/\sqrt{\sigma_2}\end{bmatrix}\begin{bmatrix}\cos(\theta_0) & \sin(\theta_0)\\ -\sin(\theta_0) & \cos(\theta_0)\end{bmatrix}\mathbf{x}.\end{aligned}$$

In the following lecture we will see how to use this in practice.