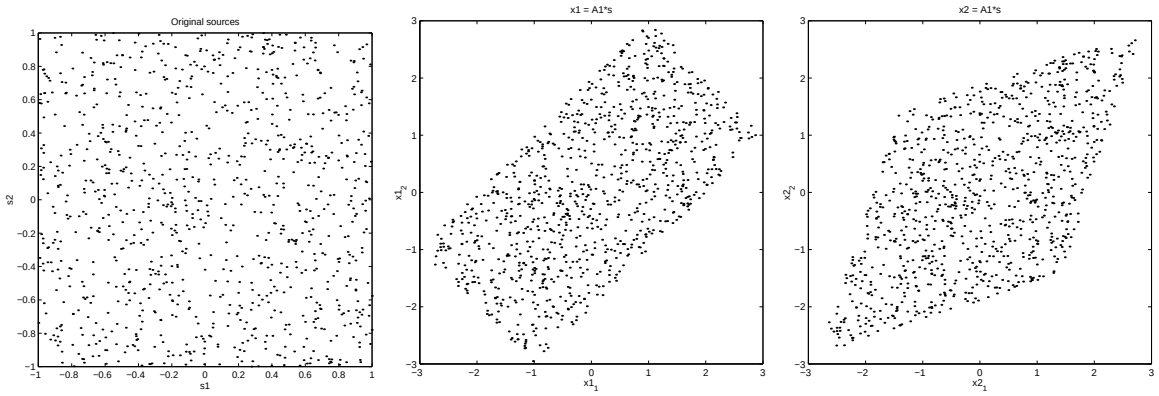
T-61.5030 Advanced course in neural computing

Solutions for exercise 6

1. (a) The original sources and the two linear mappings are



(b) The covariance is defined by $\Sigma = E\{\mathbf{x}\mathbf{x}^T\} - E\{\mathbf{x}\}E\{\mathbf{x}^T\}$. In this case the mixtures have zero mean and the second term vanishes. We have

$$\Sigma_1 = E\{\mathbf{x}_1\mathbf{x}_1^T\} = E\{\mathbf{A}_1\mathbf{s}\mathbf{s}^T\mathbf{A}_1^T\} = \mathbf{A}_1 E\{\mathbf{s}\mathbf{s}^T\}\mathbf{A}_1^T = \mathbf{A}_1\Sigma_s\mathbf{A}_1^T = \frac{1}{3}\mathbf{A}_1\mathbf{A}_1^T\,,$$

because $\Sigma_s = \mathbf{I}/3$. Similarly $\Sigma_2 = \mathbf{A}_2\mathbf{A}_2^T/3$. By computing the matrix products, we find out that

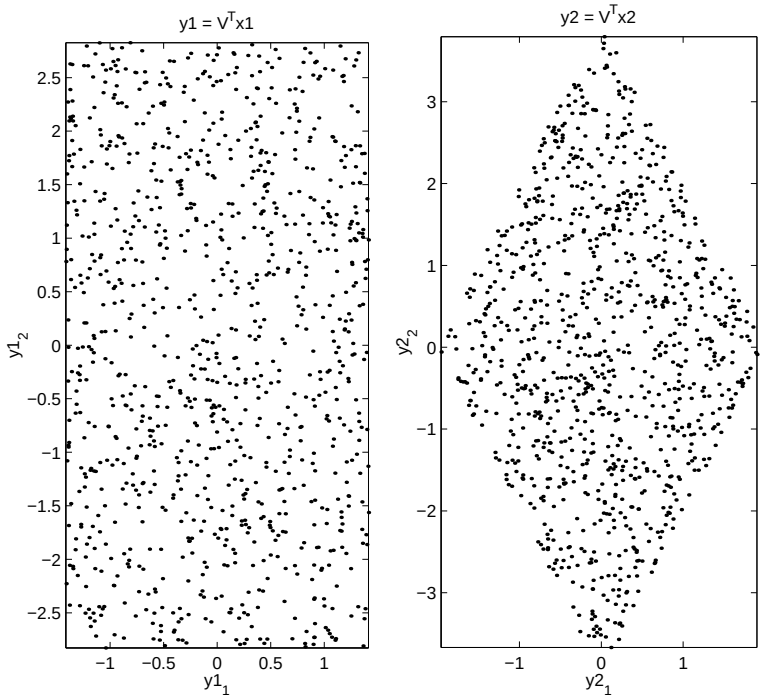
$$\Sigma_1 = \Sigma_2 = \begin{pmatrix} \frac{5}{3} & 1 \\ 1 & \frac{5}{3} \end{pmatrix}.$$

This shows that two different rotations can produce the same covariance matrix.

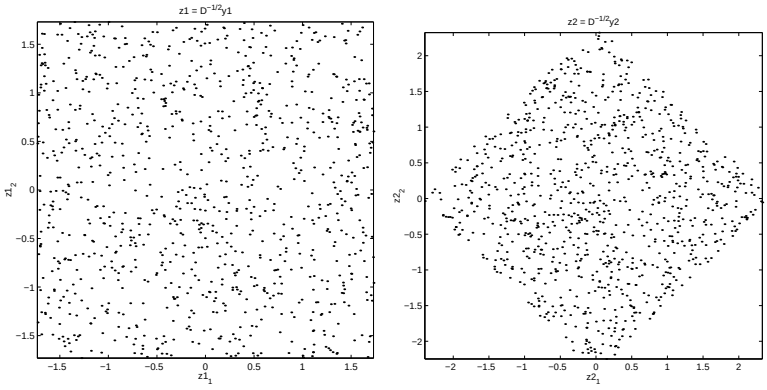
(c) The whitening can be achieved by rotating the mixtures such that the principal components are axis aligned which means that the resulting mixtures have a diagonal covariance matrix. This diagonalisation can be done by projecting the mixtures to the eigen vectors of the covariance matrix. In other words, we will find an orthogonal matrix $\mathbf{V}$ containing the eigen vectors and the corresponding diagonal matrix $\mathbf{D}$ containing the eigen values, which satisfy $\Sigma_1 = \mathbf{V}\mathbf{D}\mathbf{V}^T$. Denote the rotated mixtures by $\mathbf{y}_1 = \mathbf{V}^T\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{V}^T\mathbf{x}_2$. We then have

$$E\{\mathbf{y}_1\mathbf{y}_1^T\} = \mathbf{V}^T E\{\mathbf{x}_1\mathbf{x}_1^T\}\mathbf{V} = \mathbf{V}^T\Sigma_1\mathbf{V} = \mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{V} = \mathbf{D}\,,$$

which shows that $\mathbf{y}_1$ has a diagonal covariance matrix. The same holds for $\mathbf{y}_2$, because $\Sigma_1 = \Sigma_2$. The new coordinate systems look like this



The final step in whitening makes the covariance matrix a unit matrix. This is achieved by dividing each axis by the standard deviation in that direction. The variances are found from the diagonal of the matrix $\mathbf{D}$. We then have the whitend mixtures $\mathbf{z}_1 = \mathbf{D}^{-1/2}\mathbf{y}_1$ and $\mathbf{z}_2 = \mathbf{D}^{-1/2}\mathbf{y}_2$, where $\mathbf{D}^{-1/2}$ is defined to be the diagonal matrix where the diagonal elements of $\mathbf{D}$ have raised to the power $-1/2$. The whitened mixtures look like

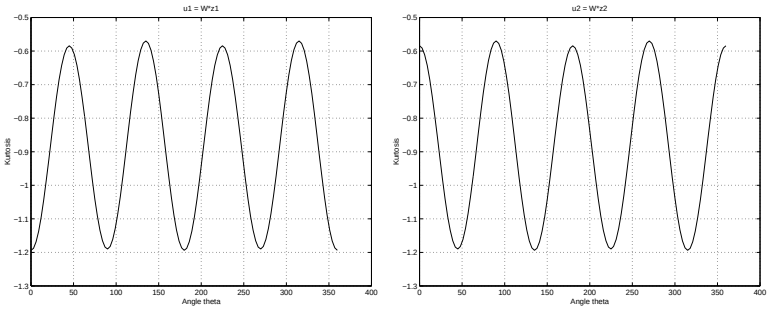


We notice that the whitening has given the original sources up to scaling and orthogonal rotation.

(d) An orthogonal rotation to $\mathbf{z}_1$ or $\mathbf{z}_2$ will leave the covariance matrix to be the unit matrix. This is because

$$E\{(\mathbf{W}\mathbf{z}_1)(\mathbf{W}\mathbf{z}_1)^T\} = \mathbf{W}E\{\mathbf{z}_1\mathbf{z}_1^T\}\mathbf{W}^T = \mathbf{W}\mathbf{I}\mathbf{W}^T = \mathbf{I}\,.$$

(e) Kurtosis can be used as a measure for non-Gaussianity. The plots for kurtosis look like

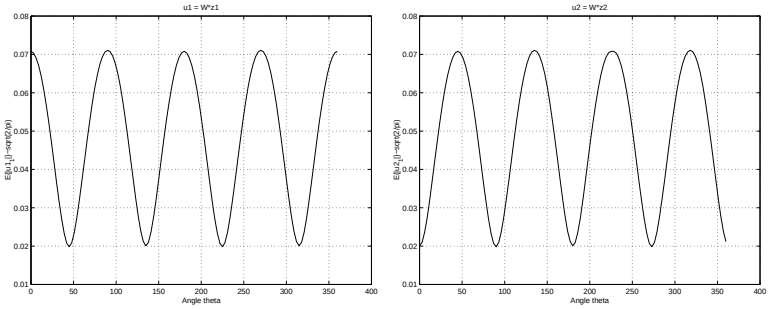


The points which are furthest away from zero, in this case the points which are most negative, are the angles where the rotation produces the least Gaussian projections. These are the ones which give the original sources. For $\mathbf{z}_1$ the rotation is 0, 90, 180 or 270 degrees while for $\mathbf{z}_2$ the rotation is 45, 135, 225 or 315 degrees.

(f) Any other measure of non-Gaussianity can be used. A simple way to construct such a measure is to take a function f and compare its expectation over the data with the expectation which would be given by a Gaussian distribution having the same variance. With $f(x) = |x|$, for example we have a measure

$$E\{|x|\} - \sqrt{2/\pi}\,,$$

because the expectation of $|x|$ for a normal distribution is $\sqrt{2/\pi}$. The plot of this measure looks like this:



In this case the values are positive and the points which are furthest from zero are the ones which are most positive.

2. Starting from the definition $I(X;Y) = h(X) - h(X|Y)$, where $h(X)$ is defined in Haykin, eq. (10.12), and $h(X|Y)$ is defined in Haykin, eq. (10.33), we have

$$I(X;Y) = -\int p(x)\ln p(x)dx + \int p(x,y)\ln p(x|y)dxdy =$$

$$-\int p(x,y)\ln p(x)dxdy + \int p(x,y)\ln\frac{p(x,y)}{p(y)}dxdy =$$

$$\int p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy = D(p(x,y)||p(x)p(y))\,.$$

3. Negentropy measures how much the (differential) entropy of a distribution differs from that of the Gaussian distribution having the same variance. If $h_G(Y_i)$ denotes the entropy of the Gaussian distribution having the same variance as the random variable Y_i, the negentropy is $J(Y_i) = h_G(Y_i) - h(Y_i)$. Recall that from all distributions with a given

variance, the Gaussian distribution has the highest entropy. Negentropy is thus always non-negative and is zero if and only if Y_i has a Gaussian distribution.

The mutual information can be written as

$$I(Y_1, \ldots, Y_n) = \sum_{i=1}^{N} h(Y_i) - h(Y_1, \ldots, Y_n)\,.$$

Therefore

$$I(Y_1, \ldots, Y_n) = \sum_{i=1}^{N} [h_G(Y_i) - J(Y_i)] - h(Y_1, \ldots, Y_n) = C - \sum_{i=1}^{N} J(Y_i)\,,$$

where C is constant. The terms $h_G(Y_i)$ are clearly constant since the variance for all Y_i was defined to be the same (unity) and the entropy of a Gaussian scalar variable depends only on its variance.

Y is defined to have a unit covariance matrix; this means that different transformation matrices $\mathbf{W}$ can differ only up to a orthogonal rotation (see the solution to exercise problem 6.1). Therefore the term $h(Y_1, \ldots, Y_n)$ is constant because entropy does not change in translations or orthogonal rotations; this is because they leave the form of the probability density untouched.

ICA can be defined as the search for the rotation which minimises the mutual information between the resulting components. The above formula for mutual information shows that this can be done also by maximising the negentropies of the components.