

Unsupervised Machine Learning Projects

Project 2

Independent component analysis

Deadline: May 8th 2015 (midnight the latest).

Exercise 1: Basics of independent component analysis (ICA)

1. Create two sets of artificial data each one consisting of 5000 points (vectors of two components) each component drawn independently from a uniform distribution (in the first set) and a Gaussian distribution (in the second set) with zero mean and unit variance in both cases. Each point in the first data set will be called \mathbf{s} , and in the second data set, \mathbf{n} . Transform these data sets using two different linear transformations

$$\mathbf{x}_1 = \mathbf{A}_1 \mathbf{s} \qquad \mathbf{x}_2 = \mathbf{A}_2 \mathbf{s} \qquad (1a)$$

$$\mathbf{y}_1 = \mathbf{A}_1 \mathbf{n} \qquad \mathbf{y}_2 = \mathbf{A}_2 \mathbf{n} \qquad (1b)$$

where the matrices \mathbf{A}_1 and \mathbf{A}_2 characterizing each linear transformation are given by

$$\mathbf{A}_1 = \begin{pmatrix} 0.4483 & -1.6730 \\ 2.1907 & -1.4836 \end{pmatrix} \qquad \mathbf{A}_2 = \begin{pmatrix} 0 & -1.7321 \\ 1.7321 & -2.0 \end{pmatrix} \qquad (2)$$

The symbols \mathbf{x}_1 , \mathbf{x}_2 , \mathbf{y}_1 and \mathbf{y}_2 are vectors of two components representing each data point after the corresponding linear transformation.

In the report: Make a figure showing the data sets before and after the linear transformations. The results for \mathbf{s} , \mathbf{x}_1 and \mathbf{x}_2 should resemble Figure 1. Describe the data before and after the linear transformations in your own words and compute the mean, covariance matrix, range, and principal components.

2. Whiten the four data sets corresponding to \mathbf{x}_1 , \mathbf{x}_2 , \mathbf{y}_1 and \mathbf{y}_2 , and describe the data after whitening (distribution, mean, covariance matrix, range, and principal components). Compare the whitened data to the original data, that is, before the linear transformations (\mathbf{s} for uniform data and \mathbf{n} for Gaussian data). Do the results look like figures 6.5 and 6.6 in the lecture notes?

In the report: Show scatter plots of the whitened data and analyse any difference with the figures in the lecture notes. Explain how the whitening matrices relate to the matrices \mathbf{A}_1 and \mathbf{A}_2 ?

3. For all four data sets, calculate the kurtosis of the projections of the whitened data in each direction. To that end, project the data onto a unit vector \mathbf{w} forming an angle α ($0 \leq \alpha \leq \pi$) with the x-axis, and compute the kurtosis of the projected data as a function of α .

In the report: Make figures with each of the four curves and show the angles for which the absolute value of the kurtosis is locally maximized. Explain the qualitative difference in the curves obtained for uniform data and Gaussian data.

4. How can you obtain estimates for \mathbf{A}_1 and \mathbf{A}_2 using the whitening matrices and the optimal projection vectors (those which maximize the absolute value of the kurtosis)?

In the report: Show your estimates for \mathbf{A}_1 and \mathbf{A}_2 in each case. Explain why finding an estimate for \mathbf{A}_1 and \mathbf{A}_2 is possible for the uniform data but not for the Gaussian data.

5. Create another data set by taking the first component of \mathbf{s} and the second component of \mathbf{n} , and apply the linear transformations mentioned in exercise 1.1 to obtained transformed sets \mathbf{z}_1 and \mathbf{z}_2 . Can you estimate \mathbf{A}_1 and \mathbf{A}_2 using the whitening matrices and the optimal projection vectors (those which maximize the absolute value of the kurtosis)?

In the report: Make figures with the new data set, show your estimations and explain your results.

Notice

The definition of kurtosis used in software implementations may differ from the definition used in the lecture notes (eq. 7.5), yielding non-zero values for Gaussian distributions instead of zero.

Hints

1. In the lecture notes, section 4.5 introduces whitening, section 5.7 the effect of scaling, and section 6.3 the relation between ICA and whitening.
2. Variability in the estimation can be assessed by resampling the data set with or without replacement and comparing the estimations from the resampled data with the original ones.

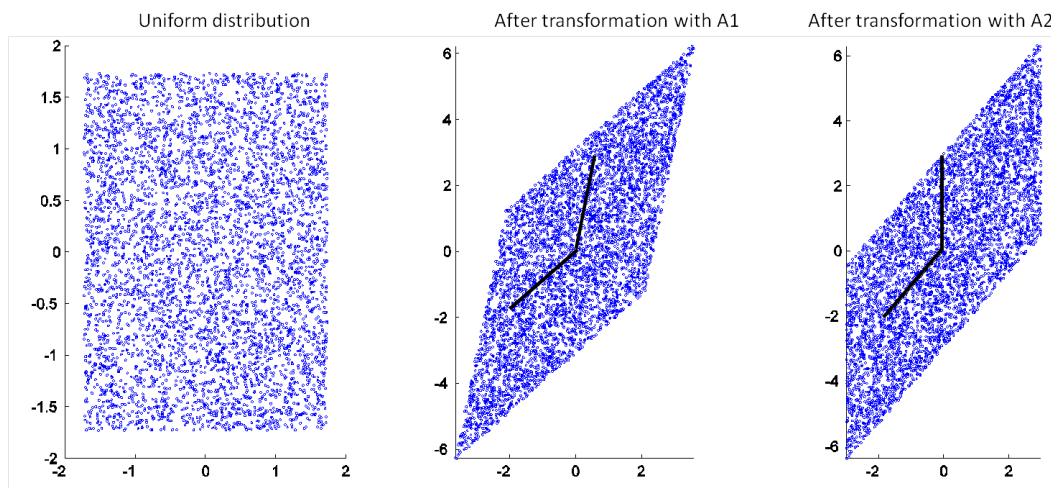


Figure 1: Scatter plots of 5000 data points uniformly distributed with zero mean and unit variance (*left*), and their linear transformation using matrices \mathbf{A}_1 (*middle*) and \mathbf{A}_2 (*right*).

Exercise 2: Kurtosis-based ICA

Let $\mathbf{s} = [s_1, \dots, s_{32}]^T$ be a random vector which consists of 32 independent random variables, each of which follow a Laplacian distribution of zero mean and unit variance (equation 7.7 in the lecture notes).

1. Generate 10000 samples of \mathbf{s} . You can use the Laplace random number generators provided by Matlab, R and Python. Use histograms with 50 bins to estimate the probability density functions (pdfs) of the following mixtures

$$y_m = \frac{\tilde{y}_m}{\sqrt{\text{Var}\tilde{y}_m}} \quad \tilde{y}_m = \sum_{i=1}^m s_i, \quad (3a)$$

for $m = 1, 2, 4, 8, 16$, and 32 (Var represents the sample variance). The variable y_m is the sum of the m first s_i , normalized to unit variance.

In the report: Show the logarithm of the six pdfs. Compare the curves to the log-distribution of a Gaussian and compute the mean square error.

2. Compute the kurtosis of y_m in function of $m \in \{1, \dots, 32\}$.
In the report: Show a figure of kurtosis as a function of m and explain the behavior of the curve. Discuss the relation between the values of the kurtosis and the log-pdfs obtained in the previous question.
3. Implement the kurtosis-based ICA algorithm in section 7.4.3 of the lecture notes. Test it on one of the data sets of Exercise 1. The algorithm should return the demixing matrix as well as the values of the objective function during the optimization. Take care to implement an appropriate stop criterion in the optimization. Test the effect of outliers in the estimation by replacing the first 10 points with vectors $[10, 10]^T$. This could represent, for example, what would happen if one does not take into consideration the border effects introduced by filtering.
In the report: Describe the algorithm together with the stopping criterion employed (less than half a page). Illustrate the effect of outliers with figures showing the result with and without them. Discuss the results (less than half a page).
4. Let $\mathbf{y} = (y_1, \dots, y_{32})^T$. The transformation $\mathbf{s} \rightarrow \mathbf{y}$ can be written as linear transformation \mathbf{A} . What is the formula to obtain \mathbf{A} ?
In the report: Derive the general formula and the one for the particular case of this project.
5. Use ICA to get an estimate $\hat{\mathbf{A}}$ of \mathbf{A} from the observations of \mathbf{y} alone. What is the average squared error $(1/32^2 \sum_{ij} (A_{ij} - \hat{A}_{ij})^2)$? How large is the error if you use 20000 samples instead of 10000? (Note: to compute the error, you must take into account that ICA delivers only results up to a permutation matrix and sign flips of the columns of $\hat{\mathbf{A}}$, see Section 6.3.3 in the lecture notes.)
In the report: Give the estimate and the error for the two sample sizes. Describe how you overcome the sign ambiguity.

Hints

1. Recall that y_m are continuous variables in the normalization of the histograms to obtain the pdf. That is, the histogram columns must be divided by the total number of counts and the bin size.

Exercise 3: Separating mixtures of images

Six different images d_i (i representing the index of the images) of size 300×300 pixels have been randomly linearly mixed obtaining six different mixtures x_j (j representing the index of the mixture images) of the same size. In other words, each pixel uv in x_j is a linear combination C_j of the pixels uv in d_1, \dots, d_6 (u and v are indexes representing the position of the pixel in the image). The mixing matrix C is therefore of size 6×6 . The mixtures have been saved in the file `mixed_images.txt` in the form of a matrix where each row represents one image. Each row was obtained by concatenating from left to right the columns of the matrix representing each image.

1. Load the data and visualize each mixture as an image. To load the image:

in Matlab/Octave, use `load mixed_images.txt`;
in R, use `x<-matrix(scan("mixed_images.txt"),6,,TRUE)`; and
in Python, use `numpy.loadtxt("mixed_images.txt")`.

In ICA, the observations (data points) are assumed identically distributed (see likelihood based ICA, section 8.1). How well do we respect this assumption in this exercise?

In the report: Show the figures with the mixtures. Discuss how to test whether samples are identically distributed or not. Do not overdo it. Just think, for example, what should happen if you divide the image in two and compute the distributions of each half. Also think of what is the distributions of natural images above and below the horizon.

2. ICA decomposes a mixture \mathbf{x} into a set of features \mathbf{s} such that $\mathbf{x} = \mathbf{A}\mathbf{s}$, where \mathbf{A} represents the mixing matrix. What can be said about the probability densities of each component of \mathbf{s} ? What is the relation between the components of \mathbf{s} and d_i ?

In the report: Write down the mathematical relation between the quantities. Discuss what can be said, thinking, for example, on the mean values, the standard deviations, and the kurtosis.

3. Whiten the data and compare it to the original images. Can you better guess what the original images look like?

In the report: Visualize the whitened data and justify your guesses in your own words (less than a third of a page). Explain the relation between the whiten mixtures and the original data.

4. Implement the ICA algorithm below:

- a) Whiten the data (the corresponding random vector is denoted by $\mathbf{z} \in \mathbb{R}^n$ below).
- b) Initialization: random for the $n \times n$ matrix \mathbf{B} , $\gamma_i = 0$ ($i = 1 \dots n$), $\mu_g = 0.8$, $\mu = 0.2$ (this are just possible values which worked fine for me).
- c) Compute $\mathbf{y} = \mathbf{B}\mathbf{z}$.
- d) Update γ_i :

$$\gamma_i \leftarrow (1 - \mu_g)\gamma_i + \mu_g \mathbf{E}(-\tanh(y_i)y_i + (1 - \tanh(y_i))^2)$$

If $\gamma_i > 0$ define $g_i(u) = -2 \tanh(u)$, else as $g_i(u) = \tanh(u) - u$

- e) Compute the objective $F = -\sum_i \gamma_i \mathbf{E}(\log \cosh(y_i))$

- f) Update \mathbf{B} by

$$\mathbf{B} \leftarrow \mathbf{B} + \mu(I + \mathbf{E}(\mathbf{g}(\mathbf{y})\mathbf{y}^T)) \mathbf{B}$$

where $\mathbf{g}(\mathbf{y}) = (g_1(y_1), \dots, g_n(y_n))^T$ and I is the identity matrix.

- g) Orthonormalize \mathbf{B} .

- h) Check convergence: If the change in F is smaller than some small threshold, return \mathbf{B} and all the values of F . Else, go back to step 4c.

This ICA algorithm is a modified version of the one in Table 8.1 in the lecture notes. Test it on the same data used to test the kurtosis-based ICA algorithm in Exercise 2 both with and without outliers (show also the values of objective F during the optimization and the values of γ_i after the optimization).

In the report: Illustrate the effect of outliers with figures showing the result with and without them. Discuss the results and compare them with those obtained in Exercise 2.

5. Apply the ICA algorithm to parse back the images.

In the report: Show the resulting images and comment on the quality of the demixing (Note: you may need to flip the signs of the obtained images). Look at the values of the learned γ_i and discuss what they imply for the six different distributions of the sources s_i .