Independent Component Analysis

Uri Shaham

October 11, 2022

1 Whitening

Let $x \in \mathbb{R}^m$ be a random vector. Whitening linearly transforms x into $\tilde{x}$, so that the coordinates of $\tilde{x}$ are uncorrelated and have unit variance, i.e., $\mathbb{E}\left[xx^T\right] = I$. Let $\mathbb{E}\left[xx^T\right] = V\Lambda V$ be the eigendecomposition of the covariance, so that $V^T x$ is the projection of x onto its principal directions. The whitening transform is given by $\tilde{x} = V\Lambda^{-\frac{1}{2}}V^T x$ (i.e., each principal component is scaled to have unit variance). Then

$$\begin{aligned}\mathbb{E}\left[\tilde{x}\tilde{x}^T\right] &= V\Lambda^{-\frac{1}{2}}V^T\mathbb{E}\left[xx^T\right]V\Lambda^{-\frac{1}{2}}V^T \\ &= V\Lambda^{-\frac{1}{2}}V^T V\Lambda V^T V\Lambda^{-\frac{1}{2}}V^T \\ &= I.\end{aligned}$$

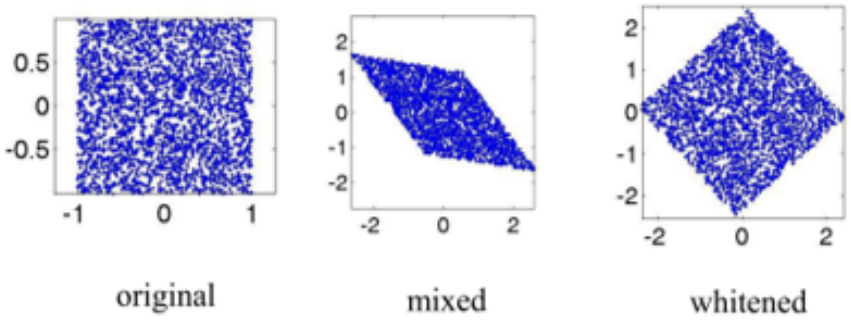

Figure 1: Example of whitening. Figure taken from `https://www.cs.cmu.edu/~bapoczos/Classes/ML10715_2015Fall/slides/ICA.pdf`

Remark 1.1. *The full procedure, with the rotation back is sometimes called ZCA whitening. People often refer to whitening transform without the rotation back, i.e., $\tilde{x} = \Lambda^{-\frac{1}{2}}V^T x$ (known as PCA whitening).*

2 Independent Component Analysis

Let $S = (S_1, \ldots S_n)^T$ be a vector of latent independent random variables (i.e., $p(s) = p(s1, \ldots, s_n) = \prod_i p(s_i)$), with zero mean and identity covariance. We observe n linear combinations of the latent

random variables, given by $x = As$, where A is unknown. Our goal is to recover S, by computing $W = A^{-1}$.

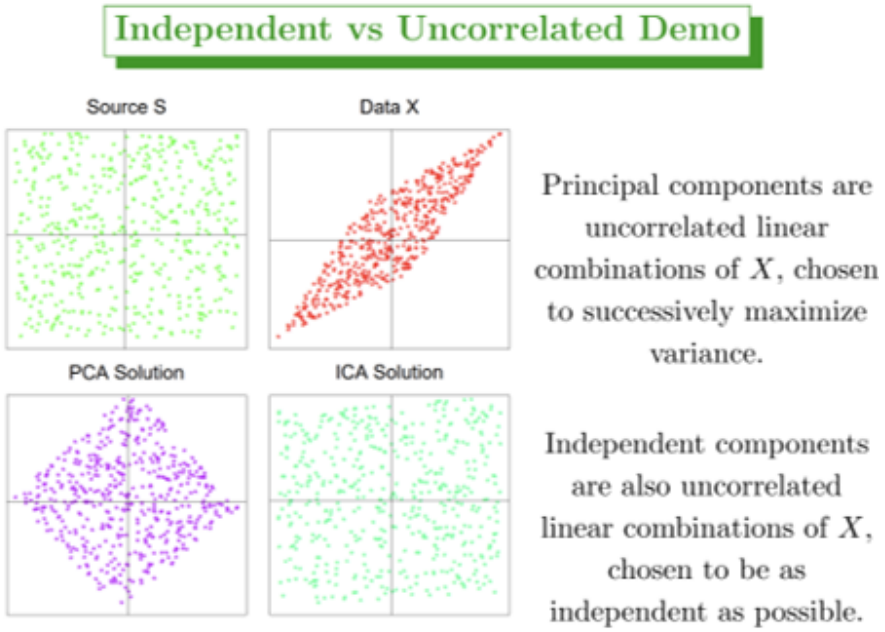

Figure 2: Difference between PCA and ICA. Figure taken from `https://hastie.su.domains/Papers/icatalk.pdf`

Suppose that $S_1, \ldots, S_n$ are all standard Gaussian. Then for any $n \times n$ rotation matrix R (i.e., $RR^T = I$) we have

$$p(RS) = \frac{1}{2\pi} \exp\left(-\frac{S^T R^T RS}{(2R^T R)^{-1}}\right) = \frac{1}{2\pi} \exp\left(-\frac{S^T S}{2I}\right),$$

which means that s cannot be recovered (or put another way, A is not identifiable). Hence from now on we assume all variables are non Gaussian.

3 Nongaussianity

CLT (Lyapunov version - not necessarily i.d): sum of independent random variables converges in distribution to normal. Thus, intuitively, a $X_j = A_j^T S$ is "more Gaussian" any of the s_i's. Thus, to find the first component, we can seek for a vector w which maximizes nonGaussianity.

3.1 Negentropy

Definition 3.1. *The differential entropy of a random variable Y with density f is $Y(y) \int f(y) \log f(y) dy$*

Fact 3.2. *A Gaussian random variable has the largest entropy among all random variables with equal variance.*

Thus, entropy can be used as a measure of nongaussianity.

Definition 3.3 (Negentropy)**.** *The Negentropy of a random variable Y is defined as $J(Y) = H(Y_{Gauss}) - H(Y)$, where $H(Y_{Gauss}) = \frac{1}{2}\log(2\pi e\sigma)$ is the entropy of a Gaussian random variable with the same variance matrix as Y*

4 Solving ICA

4.1 Whitening

Let's write $\tilde{x} = E\Lambda^{-\frac{1}{2}}E^T x = E\Lambda^{-\frac{1}{2}}E^T As = \tilde{A}s$.

Then

$$I = \mathbb{E}\left[\tilde{x}\tilde{x}^T\right] = \tilde{A}\mathbb{E}\left[ss^T\right]\tilde{A}^T = AA^T$$

, i.e., $\tilde{A}$ is orthonormal, so the problem now reduces to recovering an orthonormal matrix, which is a simpler problem for estimation (only half the degrees of freedom).

We want:

- find y (estimation of s)
- find W (estimation of A^{-1}.

Solution process:

- Remove mean (so $\mathbb{E}[x] = 0$
- Whitening: $\mathbb{E}\left[\tilde{x}\tilde{x}^T\right] = I$
- Find orthogonal W by optimizing an objective

Negentropy is typically estimated by a non-quadratic function G (e.g., $G_1(y) = y^4$, $G_2(y) = -\exp[(y^2))$, as

$$J(Y) \propto \left(\mathbb{E}[G(Y)] - \mathbb{E}[G(z)]\right)^2,$$

where z is a standard Gaussian random variable.

The minimization problem can be solved using standard methods, e.g., Newton's method

$$w^{(t+1)} = w^{(t)} - \nabla^2(w^{(t)})_{-J}^{-1}\nabla_{-J}(w^{(t)})$$

, where expectations are replaced by sample means.

For the first combination w, the requirement unit variance $\mathrm{Var}\left(W^T\tilde{x}\right) = 1$, together with the fact that $\tilde{x}$ is whitened, is equivalent to requiring that w is a unit vector. This can be implemented by rescaling w_t after each iteration of the optimization procedure.

For subsequent combination, we want each vector w to live in the orthogonal complement of the w's found so far, which we can achieve by applying Gram-Shmidt:

$$w_k \leftarrow w_k - \sum_{i=1}^{k-1} w_k^T w_i w_i.$$

Homework

1. Express whitening in terms of SVD matrices.
2. Let Y_1, Y_2 be independent random variables, and h_1, h_2 be arbitrary functions. Prove that $\mathbb{E}\left[h_1(Y_1)h_2(Y_2)\right] = \mathbb{E}\left[h_1(Y_1)\right]\mathbb{E}\left[h_2(Y2)\right]$
3. Prove that independent random variables are uncorrelated
4. Show that uncorrelated random variables are not necessarily independent.
5. Programming: (sklearn can be used)
   (a) Create two 1D time series as the latent components (where each time step is considered as a sample). Plot the recovered components using PCA and ICA.
   (b) Create a point cloud in 2d. Plot the directions found by PCA and ICA.