Independent Component Analysis

Uri Shaham

March 4, 2024

1 Whitening

Let $X \in \mathbb{R}^m$ be a zero-mean random vector. Whitening linearly transforms X into $\tilde{X}$, so that the coordinates of $\tilde{X}$ are uncorrelated and have unit variance, i.e., $\mathbb{E}\left[\tilde{X}\tilde{X}^T\right] = I$. 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, as in PCA. 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

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

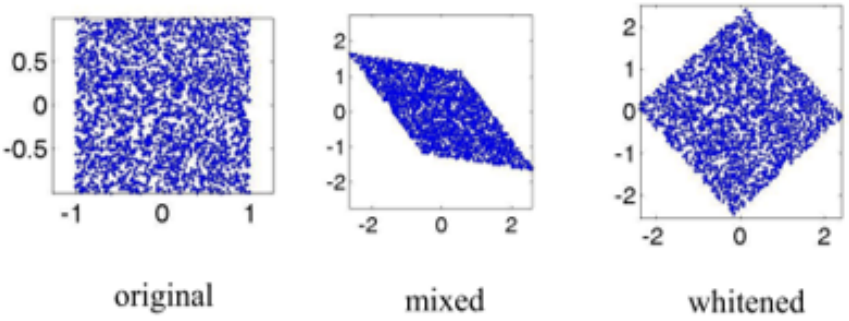

Figure 1: Example of whitening. Figure taken from `https://www.cs.cmu.edu/~bapoczos/Classes/ML10715_2015Fall/slides/ICA.pdf`

Remark 1.1. *The above procedure, with the rotation back (i.e., the leftmost multiplication by V) 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). You will show in homework that if $X_n = U\Sigma V^T$ is a $n \times d$ data matrix, PCA whitening $\Lambda^{-\frac{1}{2}}V^T X_n^T$ simply returns U^T.*

2 Independent Component Analysis

Let $S = (S_1, \dots S_n)^T$ be a vector of latent independent random variables (i.e., $\Pr(S) = \Pr(S_1, \dots, S_n) = \prod_i \Pr(S_i)$), with zero mean and identity covariance. We observe n linear combinations of the latent random variables, given by $X = AS$, where $A \in \mathbb{R}^{n \times n}$ is unknown. Our goal is to recover S, by computing $W = A^{-1}$.

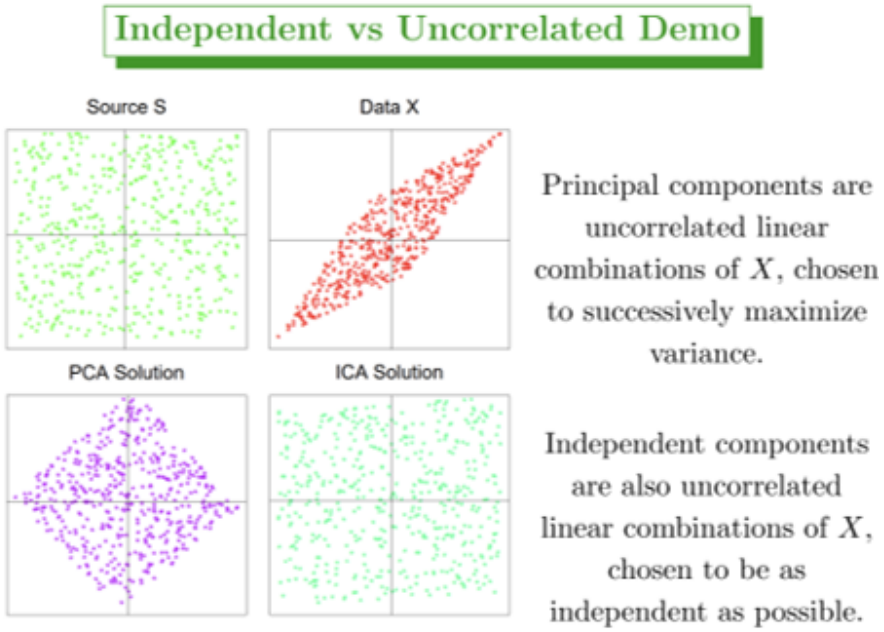

Figure 2: Difference between PCA and ICA. Figure taken from `https://hastie.su.domains/Papers/icatalk.pdf`

Suppose that $S_1, \dots, S_n$ are all standard Gaussian. Note that for every matrix B, BS is also Gaussian. Then for any $n \times n$ rotation matrix R (i.e., $RR^T = I$) we have

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

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

Lyapunov's version of the central limit theorem asserts that sum of independent (not necessarily identically distributed) random variables converges in distribution to normal. Thus, intuitively, a X_j, which is the dot product between the j'th row of A and S is "more Gaussian" any of the S_i's.

We want to recover one of the latent factors S_i, via $Y := w^T X = (w^T A)S$, which is a linear combination of the latent factors as well. Hence, to recover one of the components, we wish to find w

which maximizes the nonGaussianity of $w^T X$. A popular measure for nonGaussianity is negentropy, described next.

3.1 Negentropy

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

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

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 as Y.*

Computing $h(Y)$ is hard, as it requires a nonparametric estimation of the density $f(Y)$. Hence, one typically use approximations for it. Specifically, negentropy is typically estimated by a non-quadratic function G (e.g., $G_1(y) = y^4$, $G_2(y) = -\exp(y^2)$, as

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

where Z is a standard Gaussian random variable.

4 Solving ICA

We will aim to find an approximation Y of S. Since independent components are uncorrelated, we can restrict our search to matrices Y_n which are orthogonal, hence whitening can use as a starting point. Hence prior to the optimization, we preprocess the data matrix X_n by subtracting the mean from each column, folowwing by whitening.

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

$$w^{(t+1)} = w^{(t)} - \left(\nabla^2 \tilde{J}(\tilde{X}_n w^{(t)})\right)^{-1} \nabla \tilde{J}(\tilde{X}_n w^{(t)}),$$

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:

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

Further Reading

A good ICA tutorial is `https://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/HyvO00-icatut.pdf`.