Blind Source Separation 2.

We first need to introduce the idea of data "whitening". First, given a vector of data $\vec{x}(K)$, we can always take

$$\vec{x}(K) - E(\vec{x})$$

to transform it into zero-mean data. We'll apply this without further mention to assume that <u>all</u> our data is zero mean.

Given such a data vector $\vec{x}(K)$, we let

$$R_{xx} = E(xx^T)$$

be the covariance matrix, as usual.

If we take the SVD of R_{xx},

we get

$$R_{xx} = V_x \Lambda_x V_x^T$$

we can now apply V_x to the data to get a new time series:

$$\vec{y}(K) = V_x^T \vec{x}(K)$$

Notice that

$$\vec{y}\vec{y}^T = V_x^T \vec{x}\vec{x}^T V_x$$

so

$$E(\vec{y}\vec{y}^T) = V_x^T E(\vec{x}\vec{x}^T) V_x$$

$$= V_x^T (V_x \Lambda_x V_x^T) V_x$$

$$= \Lambda_x,$$

so the entries in $\vec{y}(K)$ are uncorrelated.

This process, reasonably enough, is called <u>decorrelation</u>. On the other hand, the entries in $\vec{y}(K)$ <u>don't</u> have unit variance. We solve that by rescaling.

Lemma. If $\vec{x}(K)$ is zero-mean and

$$R_{xx} = E(\vec{x}\,\vec{x}^T) = V_x \Lambda_x V_x^T, \text{ then}$$

$$\vec{y}(K) = \Lambda_x^{-1/2} V_x^T \vec{x}(K)$$

has $R_{yy} = E(\vec{y}\,\vec{y}^T) = I_n$.

Proof. We just write out

$$\begin{aligned}
R_{yy} &= E(\Lambda_x^{-1/2} V_x^T \vec{x}\,\vec{x}^T V_x \Lambda_x^{-1/2}) \\
&= \Lambda_x^{-1/2} V_x^T R_{xx} V_x \Lambda_x^{-1/2} \\
&= \Lambda_x^{-1/2} V_x^T (V_x \Lambda_x V_x^T) V_x \Lambda_x^{-1/2} \\
&= \Lambda_x^{-1/2} \Lambda_x \Lambda_x^{-1/2} = I_n. \qquad \square
\end{aligned}$$

We call this a "whitening" transformation because

Definition. A random vector $\vec{x}(K)$ is called a <u>white noise vector</u> if the components of $\vec{x}(K)$ are selected from <u>statistically independent probability distributions with zero mean and finite variance</u>.

Now suppose we have our usual BSS setup

$$\vec{x}(K) = H\vec{s}(K)$$

If we ~~tran~~ whiten the data, we get

$$\vec{y}(K) = Q\vec{x}(K) = QH\vec{s}(K)$$

(where $Q = \Lambda_x^{-1/2} V_x^T$, as above.)

Notice that if we take $A = QH$ as the new mixing matrix, we have

$$\vec{y}(K) = A\,\vec{s}(K),$$

so

$$R_{yy} = E(\vec{y}\,\vec{y}^T)$$

$$= E(A\,\vec{s}\,\vec{s}^T A^T)$$

$$= A\,E(\vec{s}\,\vec{s}^T)\,A^T.$$

But we assumed that $E(\vec{s}\,\vec{s}^T) = R_{ss} = I_n$ and we just assured that $R_{yy} = I_n$, so this implies

$$I_n = A A^T$$

and the new mixing matrix is <u>orthogonal</u>. Thus the <u>demixing</u> matrix W should be given by $A^T = A^{-1} = W$.

Now suppose we only want to get e of the n signals back. We'd like to express the problem in terms of minimizing the mutual information of the recovered signals.

The mutual information is given by Kullback-Leibler divergence.

The Kullback-Leibler divergence is a measure of the difference between probability measures.

If we have measures P, Q on X so that

P is absolutely continuous with respect to Q

-or-

$Q(A) = 0 \Rightarrow P(A) = 0.$

~~The~~

Then the Radon-Nikodym derivative $\frac{dP}{dQ}$ exists. ~~and~~ ~~is~~ This is a measurable function so that

$$P(A) = \int_{a \in A} \left(\frac{dP}{dQ}\right)(a)\, dQ$$

for all subsets ~~of~~ A of X.

We can then write

$$D_{KL}(P \| Q) = \int_X \left(\frac{dP}{dQ} \ln\left(\frac{dP}{dQ}\right)\right) dQ$$

which is the entropy of P with respect to Q. If Q and P are the same measure, then this is zero, otherwise it increases as P differs from Q by larger and larger amounts.

We can use this idea to measure whether two ~~distributions~~ variables are independent. Here's the idea: ~~If P and Q are independent, then on X×X~~

JH

Suppose we are given a measure μ on $X \times X$. We can generate two ~~distributions~~ measures on X by pushing forward by

$$\pi_1 : X \times X \to X, \quad \pi_1(x,y) = x$$

$$\pi_2 : X \times X \to X, \quad \pi_2(x,y) = y$$

to two new measures $(\pi_1)_* \mu$, $(\pi_2)_* \mu$. If μ is the joint distribution of independent measures on X, then

$$\mu = \underbrace{(\pi_1)_* \mu \times (\pi_2)_* \mu}_{\text{marginal distributions}},$$

so it makes sense to try to measure the "degree of independence" of x and y by computing KL-divergence from the joint distribution to the product of the marginals.

Definition. The <u>mutual information</u> of random variables (X_1, X_2) sampled according to μ on $X \times X$ is defined by

$$I(X_1, X_2) = D_{KL}\left(\mu \,\|\, (\pi_1)_*\mu \cdot (\pi_2)_*\mu\right).$$

~~We can view BSS as the problem of choosing a demixing matrix W so that the mutual (summed)~~

We can extend this definition to measure the mutual information of any number of variables by

$$I(X_1, \dots, X_k) = D_{KL}\left(\mu \,\|\, (\pi_1)_*\mu \times \dots \times (\pi_k)_*\mu\right).$$

So suppose we have (prewhitened) signals $\vec{x}(\ell)$. Amari's minimization of MI algorithm proposes that we solve

$$\text{"} \underset{W_K \in V_K(\mathbb{R}^n)}{\text{minimize}} \; I(W_K \vec{x}(\ell)) \text{"}.$$

Of course, there is a serious question here: how do we estimate the entropy of μ with respect to $(\pi_1)_* \mu \times \dots \times (\pi_K)_* \mu$ from the data?

~~We first note~~

$$\cancel{D_{KL}(P \| Q) = -E_p(\ln q(x)) + E_p(\ln p(x))}$$

Amari observes that when the number of measurements n is equal to the number of unmixed signals K ~~the~~

~~entropy in~~

~~H The KL divergence can be written as~~

if we write the joint distribution of the signals as μ (a function on $\mathbb{R}^n$) ~~and the~~ wrt Lebesgue measure on $\mathbb{R}^n$ and the marginal distributions as μ_i (the density of $(\pi_i)_* \mu$ on $\mathbb{R}$ ~~as~~ with respect to Lebesgue measure), then

$$I(\cancel{\ }\,\vec{X}) = \int_{\mathbb{R}^n} \mu \log \frac{\mu}{\prod \mu_i} \, dVol$$

$$= \int_{\mathbb{R}^n} \mu \log \mu \, dVol - \int_{\mathbb{R}^n} \mu \log \prod \mu_i \, dVol$$

$$= \int_{\mathbb{R}^n} \mu \log \mu \, dVol - \sum \int_{\mathbb{R}^n} \mu \log \mu_i \, dVol.$$

The first term is the "differential entropy" of μ. Using the change of variables formula, if we transform by another mixing matrix W, we get the differential entropy to.

$$\cancel{\int_{\mathbb{R}^n} \mu \log \mu \, dVol = \int_{\mathbb{R}^n} \mu \log}$$

change by writing the new density as $(\det W)\mu$. In this case, we get

$$\int (\det W \cdot \mu) \log(\det W \cdot \mu) \overset{(\det W)^{-1}}{d\mathrm{Vol}}$$

$$= \int \mu \log(\det W \cdot \mu)\, d\mathrm{Vol}$$

$$= \int \mu \log \mu \, d\mathrm{Vol} + \int \mu \log(\det W)\, d\mathrm{Vol}$$

$$= \int \mu \log \mu \, d\mathrm{Vol} + \log(\det W).$$

or "old entropy + $\log(\det W)$ = new entropy".

Now we have to consider

$$\int_{\mathbb{R}^n} \mu \log \mu_i \, d\mathrm{Vol}.$$

Now

$$\mu_i(x_i) = \int_{\mathbb{R}^{n-1}} \mu(x_1, \dots, x, \dots x_n)\, dx_1 \wedge \dots \wedge \widehat{dx_i} \wedge \dots \wedge dx_n$$

so we have μ_i only depending on x_i.

Thus

$$\int_{\mathbb{R}^n} \mu \log \mu_i \; dx_1 \wedge \dots \wedge dx_n =$$

$$\int_{x_i} \log \mu_i \left(\int_{\mathbb{R}^{n-1}} \mu(x_1, \dots, x, \dots, x_n)\, dx_1 \wedge \dots \wedge \widehat{dx_i} \wedge dx_n \right) dx_i$$

$$= \int_{x_i} \mu_i \log \mu_i \, dx_i,$$

which is just the (differential) entropy of the ith (demixed) signal.

We now face a challenge:

1) The joint entropy doesn't depend on demixing matrix W_n.

2) The entropy of product of marginals with respect to joint density is simply a sum of one-variable entropies.

Now we must estimate <u>these</u> entropies, from the data.