Slide11

Haykin Chapter 10:

Information-Theoretic Models

CPSC 636-600

Instructor: Yoonsuck Choe

Spring 2015

ICA section is heavily derived from Aapo Hyvärinen's ICA tutorial:
http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/.

Shannon's Information Theory

- Originally developed to help design communication systems that are efficient and reliable (Shannon, 1948).
- It is a deep mathematical theory concerned with the essence of the communication process.
- Provides a framework for: efficiency of information representation, limitations in reliable transmission of information over a communication channel.
- Gives bounds on optimum representation and transmission of signals.

Motivation

Information-theoretic models that lead to self-organization in a principled manner.

- **Maximum mutual information principle** (Linsker 1988): Synaptic connections of a multilayered neural network develop in such a way as to *maximize the amount of information preserved when signals are transformed at each processing stage of the network, subject to certain constraints*.
- **Redundancy reduction** (Attneave 1954): "Major function of perceptual machinary is to strip away some of the *redundancy* of stimulation, to describe or encode information in a form more economical than that in which it impinges on the receptors". In other words, *redundancy reduction = feature extraction*.

Information Theory Review

Topics to be covered:

- Entropy
- Mutual information
- Relative entropy
- Differential entropy of continuous random variables

Random Variables

- Notations: X random variable, x *value* of random variable.
- If X can take continuous values, theoretically it can carry infinite amount of information. However, this it is meaningless to think of infinite-precision measurement, in most cases values of X can be quantized into a finite number of discrete levels.

$$X = \{x_k | k = 0, \pm 1, ..., \pm K\}$$

- Let event $X = x_k$ occur with probability

$$p_k = P(X = x_k)$$

with the requirement

$$0 \leq p_k \leq 1, \quad \sum_{k=-K}^{K} p_k = 1$$

Uncertainty, Surprise, Information, and Entropy

- If p_k is 1 (i.e., probability of event $X = x_k$ is 1), when $X = x_k$ is observed, there is **no surprise**. You are also pretty sure about the next outcome ($X = x_k$), so you are more certain (i.e., **less uncertain**).
  - High probability events are less surprising.
  - High probability events are less uncertain.
  - Thus, surprisal/uncertainty of an event are related to the **inverse** of the probability of that event.
- You gain **information** when you go from a high-uncertainty state to a low-uncertainty state.

Entropy

- Uncertainty measure for event $X = x_k$ ($\log$ assumes $\log_2$):

$$I(x_k) = \log\left(\frac{1}{p_k}\right) = -\log p_k.$$

  - $I(x_k) = 0$ when $p_k = 1$ (no uncertainty, no surprisal).
  - $I(x_k) \geq 0$ for $0 \leq p_k \leq 1$: no negative uncertainty.
  - $I(x_k) > I(x_i)$ for $p_k < p_i$: more uncertain for less probable events.
- Average uncertainty = **Entropy** of a random variable:

$$\begin{aligned} H(X) = &\quad E[I(x_k)] \\ = &\quad \textstyle\sum_{k=-K}^{K} p_k I(x_k) \\ = &\quad -\textstyle\sum_{k=-K}^{K} p_k \log p_k \end{aligned}$$

Properties of Entropy

- The higher the $H(X)$, the higher the **potential information** you can gain through observation/measurement.
- Bounds on the entropy:

$$0 \leq H(X) \leq \log(2K+1)$$

  - $H(X) = 0$ when $p_k = 1$ and $p_j = 0$ for $j \neq k$: No uncertainty.
  - $H(X) = \log(2K+1)$ when $p_k = 1/(2K+1)$ for all k: Maximum uncertainty, when all events are equiprobable.

Properties of Entropy (cont'd)

- Max entropy when $p_k = 1/(2K+1)$ for all k follows from

$$\sum_k p_k \log\left(\frac{p_k}{q_k}\right) \geq 0$$

for two probability distributions $\{p_k\}$ and $\{q_k\}$, with the equality holding when $p_k = q_k$ for all k. (Multiply both sides with -1.)

- Kullback-Leibler divergence (relative entropy):

$$D_{p\|q} = \sum_{x \in \mathcal{X}} p_X(x) \log\left(\frac{p_X(x)}{q_X(x)}\right)$$

measures how different two probability distributions are (note that it is not symmetric, i.e., $D_{p\|q} \neq D_{q\|p}$.

Differential Entropy of Cont. Rand. Variables

- Differential entropy:

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx = -E[\log f_X(x)]$$

- Note that $H(X)$, in the limit, does not equal $h(X)$:

$$\begin{aligned}
H(X) &= -\lim_{\delta x \to 0} \sum_{k=-\infty}^{\infty} \underbrace{f_X(x_k)\delta x}_{p_k} \log(\underbrace{f_X(x)\delta x}_{p_k}) \\
&= -\lim_{\delta x \to 0} \left[\sum_{k=-\infty}^{\infty} f_X(x_k) \log(f_X(x))\delta x \right. \\
&\quad \left. + \log(\delta x) \sum_{k=-\infty}^{\infty} f_X(x_k)\delta x\right] \\
&= -\int_{-\infty}^{\infty} f_X(x_k) \log(f_X(x)) dx \\
&\quad -\lim_{\delta x \to 0} \log \delta x \int_{-\infty}^{\infty} f_X(x)\delta x \\
&= h(X) - \lim_{\delta x \to 0} \log \delta x
\end{aligned}$$

Diff. Entropy of Uniform Distribution

- Uniform distribution within interval $[0,1]$:

$f_X(x) = 1$ for $0 \leq x \leq 1$ and 0 otherwise

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty} 1 \cdot \log 1 dx \\
&= -\int_{-\infty}^{\infty} 1 \cdot 0 dx \\
&= 0.
\end{aligned} \tag{1}$$

Properties of Differential Entropy

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log|a|$

$$f_Y(y) = \frac{1}{|a|} f_Y\left(\frac{y}{a}\right)$$

$$\begin{aligned}
h(Y) = &\ -E[\log f_Y(y)] \\
= &\ -E\left[\log\left(\frac{1}{|a|} f_Y\left(\frac{y}{a}\right)\right)\right] \\
= &\ -E\left[\log f_Y\left(\frac{y}{a}\right)\right] + \log|a|.
\end{aligned}$$

Plugging in $Y = aX$ to the above, we get the desired result.

- For vector random variable $\mathbf{X}$,

$$h(\mathbf{AX}) = h(\mathbf{X}) + \log|\det(\mathbf{A})|.$$

Maximum Entropy Principle

- When choosing a probability model given a set of known states of a stochastic system and constraints, there could be potentially an infinite number of choices. Which one to choose?
- Jaynes (1957) proposed the maximum entropy principle:
  - Pick the probability distribution that maximizes the entropy, subject to constraints on the distribution.

One Dimensional Gaussian Dist.

- Stating the problem in an constrained optimization framework, we can get interesting general results.
- For a given variance σ^2, the Gaussian random variable has the largest differential entropy attainable by any random variable.
- The entropy of a Gaussian random variable X is uniquely determined by the variance of X.

Mutual Information

- **Conditional entropy**: What is the entropy in X after observing Y? How much uncertainty remains in X after observing Y?

$$H(X|Y) = H(X,Y) - H(Y)$$

where the joint-entropy is defined as

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

- **Mutual information**: How much uncertainty is reduced in X when we observe Y? The amount of reduced uncertainty is equal to the amount of information we gained!

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Mutual Information for Continuous Random Variables

- In analogy with the discrete case:

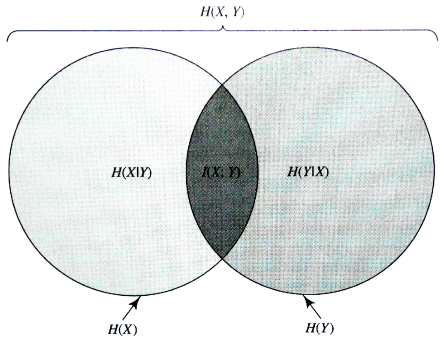
$$I(X;Y) = \int_{\infty}^{\infty}\int_{\infty}^{\infty} f_{X,Y}(x,y)\log\left(\frac{f_X(x|y)}{f_X(x)}\right)dxdy$$

- And it has the same property

$$\begin{aligned} I(X;Y) = &\quad h(X) - h(X|Y) \\ = &\quad h(Y) - h(Y|X) \end{aligned}$$

$$= h(X) + h(Y) - h(X,Y)$$

Summary



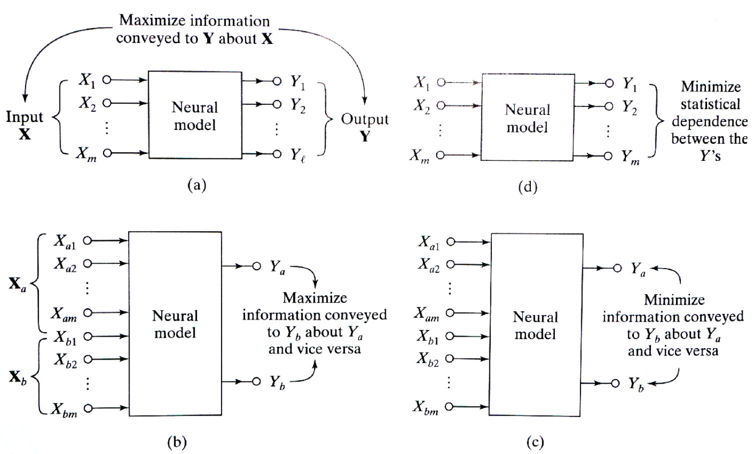
- Various relationships among entropy, conditional entropy, joint entropy, and mutual information can be summarized as shown above.

Properties of KL Divergence

- It is always positive or zero. Zero, when there is a perfect match between the two distributions.
- It is invariant w.r.t.
  - Permutation of the order in which the components of the vector random variable $\mathbf{x}$ are arranged.
  - Amplitude scaling.
  - Monotonic nonlinear transformation.
- It is related to mutual information:

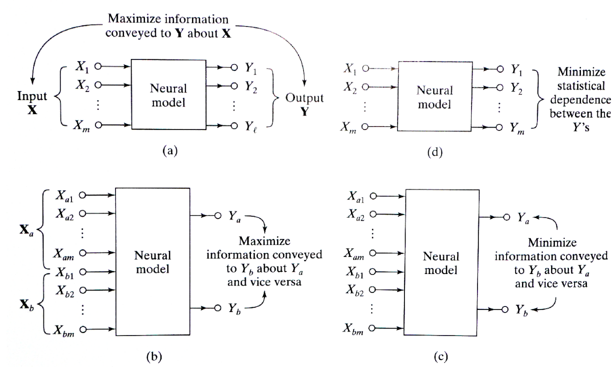
$$I(\mathbf{X};\mathbf{Y}) = D_{f_{\mathbf{X},\mathbf{Y}} \| f_{\mathbf{X}} f_{\mathbf{Y}}}$$

Application of Information Theory to Neural Network Learning



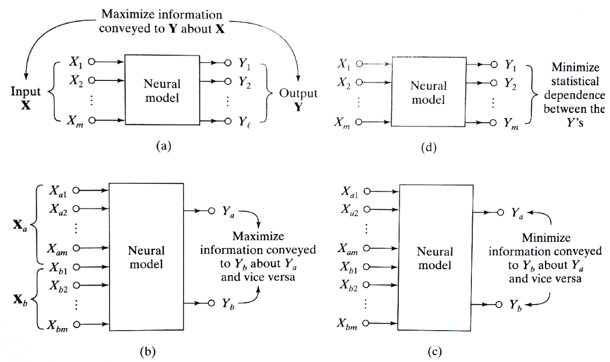
- We can use mutual information as an objective function to be optimized when developing learning rules for neural networks.

Mutual Information as an Objective Function



- (a) Maximize mutual info between input vector $\mathbf{X}$ and output vector $\mathbf{Y}$.
- (b) Maximize mutual info between Y_a and Y_b driven by near-by input vectors $\mathbf{X}_a$ and $\mathbf{X}_b$ from a *single* image.

Mutual Info. as an Objective Function (cont'd)



- (c) Minimize information between Y_a and Y_b driven by input vectors from *different* images.
- (d) Minimize statistical dependence between Y_i's.

Maximum Mutual Information Principle

- **Infomax** principle by Linsker (1987, 1988, 1989): Maximize $I(\mathbf{Y};\mathbf{X})$ for input vector $\mathbf{X}$ and output vector $\mathbf{Y}$.
- Appealing as the basis for statistical signal processing.
- Infomax provides a mathematical framework for *self-organization*.
- Relation to *channel capacity*, which defines the Shannon limit on the rate of information transmission through a communication channel.

Example: Single Neuron + Output Noise

- Single neuron with additive output noise:

$$Y = \left(\sum_{i=1}^{m} w_i X_i\right) + N,$$

where Y is the output, w_i the weight, X_i the input, and N the processing noise.

- Assumptions:
  - Output Y is a Gaussian r.v. with variance σ_Y^2.
  - Noise N is also a Gaussian r.v. with $\mu = 0$ and variance σ_N^2.
  - Input and noise are uncorrelated: $E[X_i N] = 0$ for all i.

Ex.: Single Neuron + Output Noise (cont'd)

- Mutual information between input and output:

$$I(Y;\mathbf{X}) = h(Y) - h(Y|\mathbf{X}).$$

- Since $P(Y|\mathbf{X}) = c + P(N)$, where c is a constant,

$$h(Y|\mathbf{X}) = h(N).$$

Given $\mathbf{X}$, what remains in Y is just noise N. So, we get

$$I(Y;\mathbf{X}) = h(Y) - h(N).$$

Entropy of the Normal Distribution

$$\begin{aligned}
p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
H(X) &= -\int_{-\infty}^{\infty} p(x)\log p(x)dx \\
&= -\int_{-\infty}^{\infty} p(x)\log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right]dx \\
&= -\int_{-\infty}^{\infty} p(x)\left[-\log\sqrt{2\pi\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2}\right]dx \\
&= \frac{1}{2}\log 2\pi\sigma^2 \int_{-\infty}^{\infty} p(x)dx + \frac{1}{2\sigma^2}\int_{-\infty}^{\infty} p(x)(x-\mu)^2 dx \\
&= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2} = \frac{1}{2}\left(\log 2\pi\sigma^2 + 1\right). \qquad (2)
\end{aligned}$$

Ex.: Single Neuron + Output Noise (cont'd)

- Since both Y and N are Gaussian,

$$h(Y) = \frac{1}{2}[1 + \log(2\pi\sigma_Y^2)]$$

$$h(N) = \frac{1}{2}[1 + \log(2\pi\sigma_N^2)]$$

- So, finally we get:

$$I(Y;\mathbf{X}) = \frac{1}{2}\log\left(\frac{\sigma_Y^2}{\sigma_N^2}\right).$$

- The ratio σ_Y^2/σ_N^2 can be viewed as a signal-to-noise ratio. If noise variance σ_N^2 is fixed, the mutual information $I(Y;\mathbf{X})$ can be maximized simply by *maximizing the output variance* σ_Y^2!

Example: Single Neuron + Input Noise

- Single neuron, with noise on each input line:

$$Y = \sum_{i=1}^{m} w_i(X_i + N_i).$$

- We can decompose the above to

$$Y = \sum_{i=1}^{m} w_i X_i + \underbrace{\sum_{i=1}^{m} w_i N_i}_{\text{call this } N'}$$

- N' is also a Gaussian distribution, with variance:

$$\sigma_{N'}^2 = \sum_{i=1}^{m} w_i^2 \sigma_N^2.$$

Example: Single Neuron + Input Noise

- As before:

$$h(Y|\mathbf{X}) = h(N') = \frac{1}{2}(1+2\pi\sigma_{N'}^2) = \frac{1}{2}\left[1 + 2\pi\sigma_N^2 \sum_{i=1}^{m} w_i^2\right].$$

- Again, we can get the mutual information as:

$$I(Y;\mathbf{X}) = h(Y) - h(N') = \frac{1}{2}\log\left(\frac{\sigma_Y^2}{\sigma_N^2 \sum_{i=1}^{m} w_i^2}\right)$$

- Now, with fixed σ_N^2, information is maximized by maximizing the ratio $\sigma_Y^2 / \sum_{i=1}^{m} w_i^2$, where σ_Y^2 is a function of w_i.

Lessons Learned

- Application of Infomax principle is problem-dependent.
- When $\sum_{i=1}^{m} w_i^2 = 1$, then the two additive noise models behave similarly.
- Assumptions such as Gaussianity need to be justified (it's hard to calculate mutual information without such tricks).
- Adpoting a Gaussian noise model, we can invoke a "surrogate" mutual information computed relatively easily.

Noiseless Network

- Noiseless network that transforms a random vector $\mathbf{X}$ of arbitrary distribution to a new random vector $\mathbf{Y}$ of different distribution: $\mathbf{Y} = \mathbf{W}\mathbf{X}$.
- Mutual information in this case is:

$$I(\mathbf{Y}; \mathbf{X}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}).$$

With noiseless mapping, $H(\mathbf{Y}|\mathbf{X})$ attains the lowest value ($-\infty$).

- However, we can consider the gradient instead:

$$\frac{\partial I(\mathbf{Y}; \mathbf{X})}{\partial \mathbf{W}} = \frac{\partial H(\mathbf{Y})}{\partial \mathbf{W}}.$$

Since $H(\mathbf{Y}|\mathbf{X})$ is independent of $\mathbf{W}$, it drops out.

- Maximizing mutual information between input and output is equivalent ot *maximing entropy* in the output, both with respect to the weight matrix $\mathbf{W}$ (Bell and Sejnowski 1995).

Infomax and Redundancy Reduction

- In Shannon's framework, Order and structure = Redundancy.
- *Increase* in the above *reduces* uncertainty.
- More redundancy in the signal implies less information conveyed.
- More information conveyed means less redundancy.
- Thus, Infomax principle leads to reduced reduncancy in output $\mathbf{Y}$ compared to input $\mathbf{X}$.
- When noise is present:
  - Input noise: add redundancy in input to combat noise.
  - Output noise: add more output components to combat noise.
  - *High level of noise favors redundancy of representation.*
  - *Low level of noise favors diversity of representation.*

Modeling of a Perceptual System

- Importance of redundancy in sensory messages: Attneave (1954), Barlow (1959).
- Redundancy provides *knowledge* that enables the brain to build "cognitive maps" or "working models" of the environment (Barlow 1989).
- Reduncany reduction: specific form of *Barlow's hypothesis* – early processing is to turn highly redundant sensory input into more efficient *factorial code*. Outputs become *statistically independent*.
- Atick and Redlich (1990): *principle of minumum redundancy*.

Principle of Minimum Redundancy

- Sensory signal $\mathbf{S}$, Noisy input $\mathbf{X}$, Recoding system $\mathbf{A}$, noisy output $\mathbf{Y}$.

$$\mathbf{X} = \mathbf{S} + \mathbf{N}_1$$

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}_2$$

- Retinal input includes redundant information. Purpose of retinal coding is to reduce/eliminate the redundant bits of data due to correlations and noise, before sending the signal along the optic nerve.
- *Redundancy measure* (with channel capacity $C(\cdot)$):

$$R = 1 - \frac{I(\mathbf{Y};\mathbf{S})}{C(\mathbf{Y})}$$

Principle of Minimum Redundancy (cont'd)

- Objective: find recoder matrix $\mathbf{A}$ such that

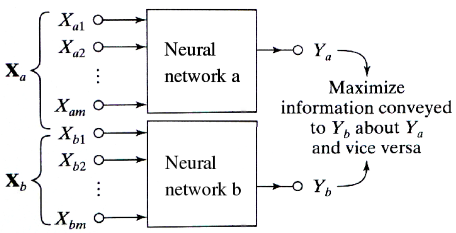
$$R = 1 - \frac{I(\mathbf{Y};\mathbf{S})}{C(\mathbf{Y})}$$

is minimized, subject to the *no information loss* constaraint:

$$I(\mathbf{Y};\mathbf{X}) = I(\mathbf{X};\mathbf{X}) - \epsilon.$$

- When $\mathbf{S}$ and $\mathbf{Y}$ have the same dimensionality and there is no noise, principle of minimum redundancy is equivalent to the Infomax principle.
- Thus, Infomax on input/output lead to reduncancy reduction.

Spatially Coherent Features



- Infomax for unsupervised processing of the image of natural scenes (Becker and Hinton, 1992).
- Goal: design a self-organizing system that is capable of learning to encode complex scene information in a simpler form.
- Objective: extract *higher-order features* that exhibit *simple coherence across space* so that representation for one spatial region can be used to produce that of representation of neighboring regions.

Spatially Coherent Features (cont'd)

- Let S denote a signal component common to both Y_a and Y_b. We can then express the outputs in terms of S and some noise:

$$Y_a = S + N_a$$

$$Y_b = S + N_b$$

and further assume that N_a and N_b are independent and zero-mean Gaussian. Also assume S is Gaussian.

- The mutual information then becomes

$$I(Y_a;Y_b) = h(Y_a) + h(Y_b) - h(Y_a, Y_b).$$

Spatially Coherent Features (cont'd)

- With $I(Y_a;Y_b) = h(Y_a) + h(Y_b) - h(Y_a, Y_b)$ and

$$h(Y_a) = \frac{1}{2}\left[1 + \log\left(2\pi\sigma_a^2\right)\right]$$

$$h(Y_b) = \frac{1}{2}\left[1 + \log\left(2\pi\sigma_b^2\right)\right]$$

$$h(Y_a, Y_b) = 1 + \log(2\pi) + \frac{1}{2}\log|\det(\Sigma),|$$

$$\Sigma = \begin{bmatrix} \sigma_a^2 & \rho_{ab}\sigma_a\sigma_b \\ \rho_{ab}\sigma_a\sigma_b & \sigma_b^2 \end{bmatrix} \quad \text{(covariance matrix)}$$

$$\rho_{ab} = \frac{E[(Y_a - E[Y_a])(Y_b - E[Y_b])]}{\sigma_a\sigma_b} \quad \text{(correlation)}$$

we get

$$I(Y_a;Y_b) = -\frac{1}{2}\log\left(1 - \rho_{ab}^2\right).$$

Spatially Coherent Features (cont'd)

- The final results was:

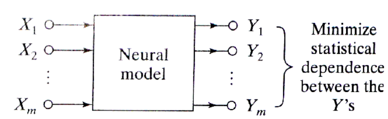
$$I(Y_a;Y_b) = -\frac{1}{2}\log\left(1 - \rho_{ab}^2\right).$$

- That is, maximizing information is equivalent to maximizing *correlation* between Y_a and Y_b, which is intuitively appealing.
- Relation to *canonical correlation* in statistics:
  - Given random input vectors $\mathbf{X}_a$ and $\mathbf{X}_b$,
  - find two weight vectors $\mathbf{w}_a$ and $\mathbf{w}_b$ so that
  - $Y_a = \mathbf{w}_a^T\mathbf{X}_a$ and $Y_b = \mathbf{w}_b^T\mathbf{X}_b$ have **maximum correlation** between them (Anderson 1984).
  - Applications: stereo disparity extraction (Becker and Hinton, 1992).

Spatially Coherent Features

- When the inputs come from two separate regions, we want to *minimize* the mutual information between the two outputs (Ukrainec and Haykin, 1992, 1996).
- Applications include when input sources such as different polarizations of the signal are imaged: mutual information between outputs driven by two orthogonal polarizations should be minimized.

Independent Components Analysis (ICA)



- Unknown random source vector $\mathbf{U}(n)$:

$$\mathbf{U} = [U_1, U_2, ..., U_m]^T,$$

where the m components are supplied by a set of *independent sources*. Note that we need a series of source vectors.

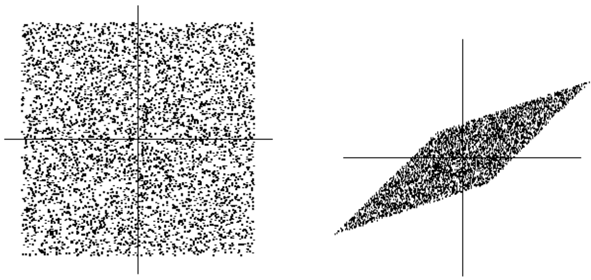
- $\mathbf{U}$ is transformed by an unknown *mixing matrix* $\mathbf{A}$:

$$\mathbf{X} = \mathbf{A}\mathbf{U},$$

where

$$\mathbf{X} = [X_1, X_2, ..., X_m]^T.$$

ICA (cont'd)

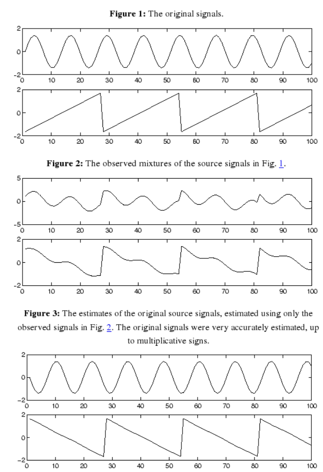


$$A = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}.$$

- Left: u_1 on x-axis, u_2 on y-axis (source)
- Right: x_1 on x-axis, x_2 on y-axis (observation)
- Thoughts: how would PCA transform this?

Examples from Aapo Hyvarinen's ICA tutorial:
`http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/`.

ICA (cont'd)



Examples from AApo Hyvarinen's ICA tutorial:
`http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/`.

ICA (cont'd)

- In $\mathbf{X} = \mathbf{AU}$, both $\mathbf{A}$ and $\mathbf{U}$ are **unknown**.
- **Task**: find an estimate of the *inverse* of the mixing matrix (the **demixing matrix** $\mathbf{W}$)

  $$\mathbf{Y} = \mathbf{WX}.$$

  The hope is to recover the unknown source $\mathbf{U}$. (A good example is the *cocktail party problem*.)

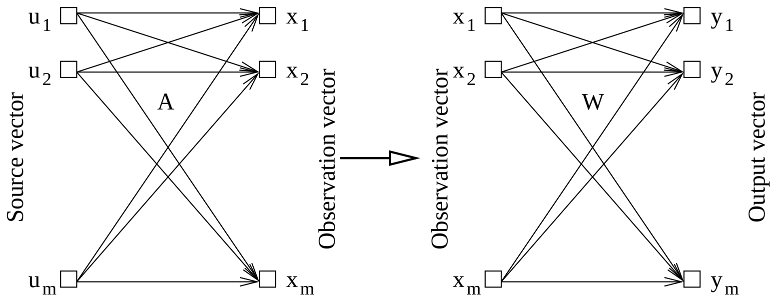
  This is known as the **blind source separation** problem.
- **Solution**: It is actually feasible, but certain ambiguities cannot be resolved: sign, permutation, scaling (variance). Solution can be obtained by enforcing **independence** among components of $\mathbf{Y}$ while adjusting $\mathbf{W}$, thus the name *independent components analysis*.

ICA: Ambiguities

Consider $\mathbf{X} = \mathbf{AU}$, and $\mathbf{Y} = \mathbf{WX}$.

- Permutation: $\mathbf{X} = \mathbf{AP}^{-1}\mathbf{PU}$, where $\mathbf{P}$ is a permutation matrix. Permuting $\mathbf{U}$ and $\mathbf{A}$ in the same way will give the same $\mathbf{X}$.
- Sign: the model is unaffected by multiplication of one of the sources by -1.
- Scaling (variance): estimate scaling up $\mathbf{U}$ and scaling down $\mathbf{A}$ will give the same $\mathbf{X}$.

ICA: Neural Network View



- The mixer on the left is an *unknown* physical process.
- The demixer on the right could be seen as a neural network.

ICA: Independence

- Two random variables X and Y are *statistically independent* when

$$f_{X,Y}(x,y) = f_X(x) f_Y(y),$$

where $f(\cdot)$ is the probability density function.

- A weaker form of independence is *uncorrelatedness* (zero covariance), which is

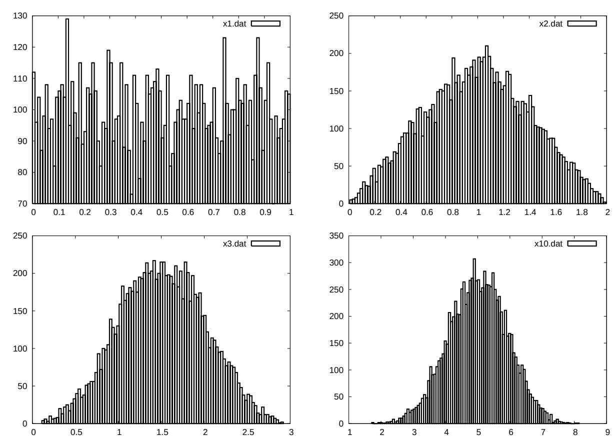
$$E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y] = 0,$$

i.e.,

$$E[XY] = E[X]E[Y].$$

- Gaussians are bad: When the unknown source is Gaussian, any orthogonal transformation A results in the same Gaussian distribution.

Statistical Aside: Central Limit Theorem



- When i.i.d. random variables $X_1, X_2, \ldots$ are added to get another random variable X, X tends to a normal distribution.
- So, Gaussians are prevalent and hard to avoid in statistics.

ICA: Non-Gaussianity

- Non-Gaussianity can be used as a measure of independence.
- The intuition is as follows:

$$\mathbf{X} = \mathbf{AU}, \quad \mathbf{Y} = \mathbf{WX}$$

Consider one component of $\mathbf{Y}$:

$$Y_i = [W_{i1}, W_{i2}, \ldots, W_{im}]\mathbf{X}$$

$$Y_i = \underbrace{[W_{i1}, W_{i2}, \ldots, W_{im}]\mathbf{A}}_{\text{call this } \mathbf{Z}^T}\mathbf{U}$$

So, Y_i is a linear combination of random variables U_k ($Y_i = \sum_{j=1}^{m} Z_i U_i$), so it is more Gaussian than any individual U_k's. The Gaussianity is *minimized* when Y_i equals one of U_k's (one Z_p is 1 and all the rest 0).

ICA: Measures of Non-Gaussianity

There are several measures of non-Gaussianity

- Kurtosis
- Negentropy
- etc.

ICA: Kurtosis

- Kurtosis is the fourth-order cumulant.

$$\text{Kurtosis}(Y) = E[Y^4] - 3\left(E\left[Y^2\right]\right)^2.$$

- Gaussian distributions have kurtosis = 0.
- More peaked distributions have kurtosis $>$ 0.
- More flatter distributions have kurtosis $<$ 0.
- **Learning**: Start with random $\mathbf{W}$. Adjust $\mathbf{W}$ and measure change in kurtosis. We can also use gradient-based methods.
- **Drawback**: Kurtosis is sensitive to outliers, and thus not robust.

ICA: Negentropy

- Negentropy J is defined as

$$J(\mathbf{Y}) = H(\mathbf{Y}_{\text{gauss}}) - H(\mathbf{Y})$$

  where Y_{gauss} is a Gaussian random variable that has the same covariance matrix as $\mathbf{Y}$.
- Negentropy is always non-negative, and it is zero iff $\mathbf{Y}$ is Gaussian.
- Thus, maximizing negentropy is to maximize non-Gaussianity.
- Problem is that estimating negentropy is difficult, and requires the knowledge of the pdfs.

ICA: Approximation of Negentropy

- Classical method:

$$J(Y) \approx \frac{1}{2}E[Y^3]^2 + \frac{1}{48}\text{Kurtosis}(Y)^2$$

  but it is not robust due to the involvement of the kurtoris.
- Another variant:

$$J(Y) \approx \sum_{k=1}^{p} k_i \left(E[G_i(Y)] - E[G_i(N)]\right)^2$$

  where k_i's are coefficients, $G_i(\cdot)$'s are nonquadratic functions, and N is a zero-mean, unit-variance Gaussian r.v.
- This can be further simplified by

$$J(Y) \approx \left(E[G(Y)] - E[G(N)]\right)^2$$

$$G_1(Y) = \frac{1}{a_1}\log\cosh a_1 Y, \quad G_2(Y) = -\exp(-Y^2/2).$$

ICA: Minimizing Mutual Information

- We can also aim to minimize mutual information between Y_i's.
- This turns out to be equivalent to maximizing negentropy (when Y_i's have unit variance).

$$I(Y_1; Y_2; ...; Y_m) = C - \sum_i J(Y_i)$$

where C is a constant that does not depend on the weight matrix $\mathbf{W}$.

ICA: Achieving Independence

- Given output vector $\mathbf{Y}$, we want Y_i and Y_j to be statistically independent.
- This can achieved when $I(Y_i; Y_j) = 0$.
- Another alternative is to make the probability density $f_{\mathbf{Y}}(\mathbf{y}, \mathbf{W})$ parameterized by the matrix $\mathbf{W}$ to approach the *factorial distribution*:

$$\widetilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{W}) = \prod_{i=1}^{m} \widetilde{f}_{Y_i}(y_i, \mathbf{W}),$$

where $\widetilde{f}_{Y_i}(y_i, \mathbf{W})$ is the *marginal probability density* of Y_i. This can be measured by $D_{f\|\widetilde{f}}(\mathbf{W})$.

ICA: KL Divergence with Factorial Dist

- The KL divergence can be shown to be:

$$D_{f\|\widetilde{f}}(\mathbf{W}) = -h(\mathbf{Y}) + \sum_{i=1}^{m} \widetilde{h}(Y_i).$$

- Next, we need to calculate the output entropy:

$$h(\mathbf{Y}) = h(\mathbf{W}\mathbf{X}) = h(\mathbf{X}) + \log|\det(\mathbf{W})|.$$

- Finally, we need to calculate the marginal entropy $\widetilde{h}(Y_i)$, which gets tricky. This calculation involves a polynomial activation function $\varphi(y_i)$. See the textbook for details.

ICA: Learning $\mathbf{W}$

- Learning objective is to minimize the KL divergence $D_{f\|\widetilde{f}}$.
- We can do *gradient descent*:

$$\begin{aligned} \Delta w_{ik} &= -\eta \frac{\partial}{\partial w_{ik}} D_{f\|\widetilde{f}} \\ &= \eta \left((\mathbf{W}^{-T})_{ik} - \varphi(y_i) x_k \right). \end{aligned}$$

- The final learning rule, in matrix form, is:

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \eta(n) \left[\mathbf{I} - \boldsymbol{\varphi}(\mathbf{y}(n)) \mathbf{y}^T(n) \right] \mathbf{W}^{-T}(n).$$

ICA Examples

- Visit the url `http://www.cis.hut.fi/projects/compneuro/whatisica.html` for interesting results.