Independent Component Analysis

Mixture Data

- Data that are mingled from multiple sources
  - May not know how many sources
  - May not know the mixing mechanism
- Good Representation
  - Uncorrelated, information-bearing components
    - PCA and Fisher's linear discriminant
  - De-mixing or separation
    - ICA (Independent component analysis)
- How do they differ?

PCA vs. ICA

- Independent events vs. Uncorrelated events

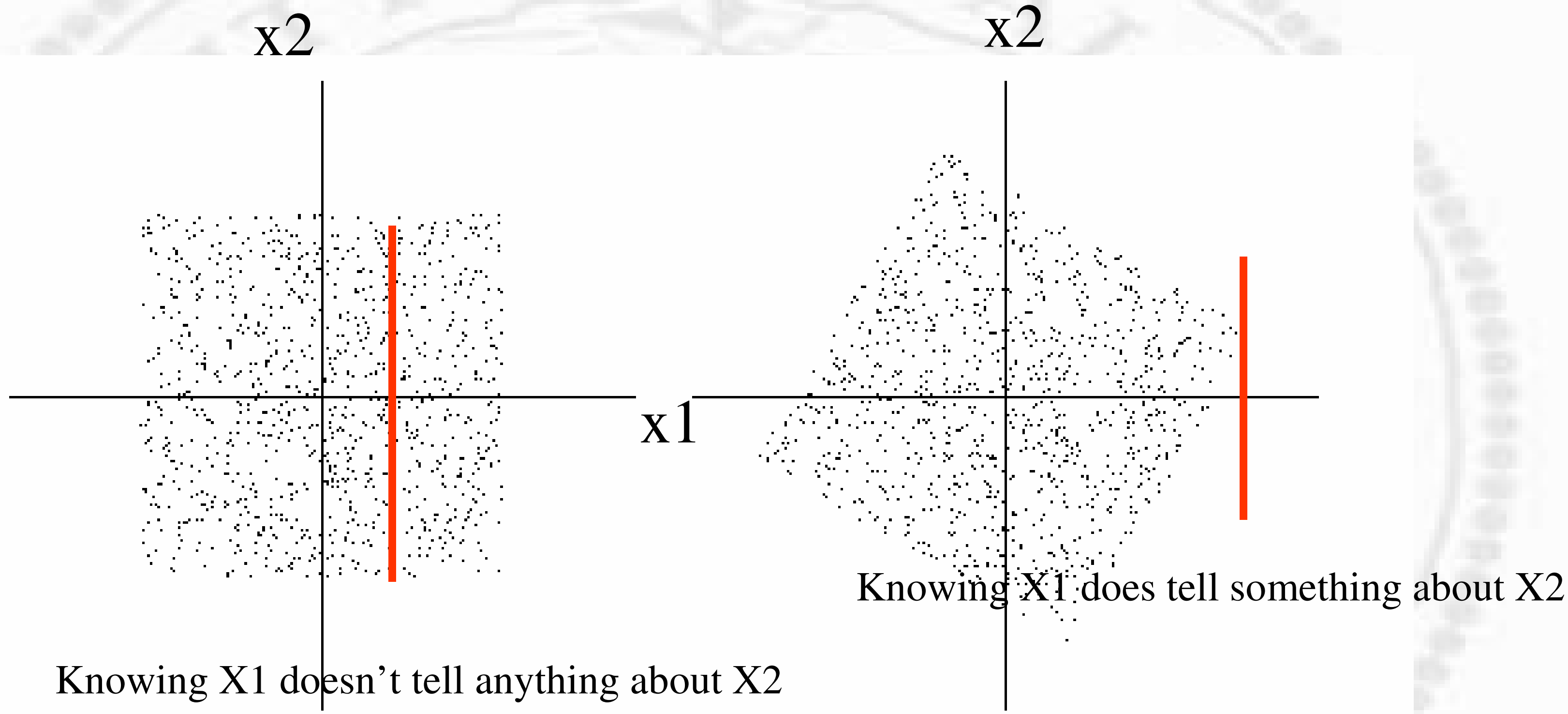


Fig. 1.4 A sample of independent components s_1 and s_2 with uniform distributions. Horizontal axis: s_1; vertical axis: s_2.

Fig. 1.5 Uncorrelated mixtures x_1 and x_2. Horizontal axis: x_1; vertical axis: x_2.

Uncorrelated vs. Independence

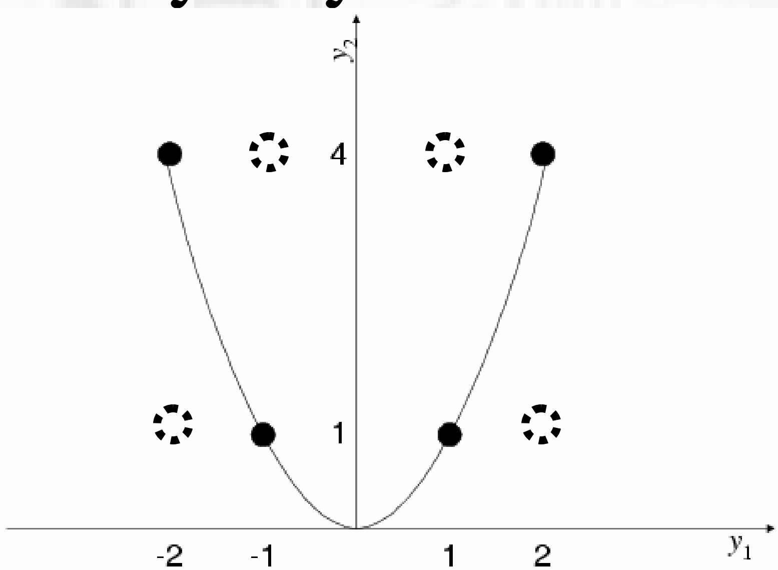
- Uncorrelated
  - Global property
  - Not valid under nonlinear transform
  - PCA requires uncorrelation
- Independence
  - Local property
  - Valid for nonlinear transform
  - ICA assumes independence

$$independence: E(g_1(x_1), g_2(x_2), \cdots, g_n(x_n)) = E(g_1(x_1)) \cdots E(g_n(x_n)) \forall g$$

$$uncorrelated: E((x_1 - Ex_1)(x_2 - Ex_2)) = 0$$

Uncorrelated vs. Independence

- Independence is stronger, requiring *every possible function* of x1 to be uncorrelated with x2
- E((y1-E(y1))(y2-E(y2))=0 -> uncorrelated
- y2= $y1^2$ -> not independent



		y_1				
		−2	−1	1	2	
y_2	1	0.00	0.25	0.25	0.00	0.50
	4	0.25	0.00	0.00	0.25	0.50
		0.25	0.25	0.25	0.25	1.00

Uncorrelated vs. Independence

- Discrete variables X1 and X2
- (0,1), (0,-1),(1,0),(-1,0) all with ¼ probability
- X1 and X2 are uncorrelated
- $E(x1^2 x2^2)=0 != 1/4 = E(x1^2)E(x2^2)$

ICA Limitation

- Any symmetrical distribution of x1 and x2 around origin (centered at Ex1 and Ex2) is uncorrelated
- Corollary: ICA does not apply to Gaussian variables
  - Because any orthogonal transform (rotation and reflection) of Gaussian doesn't change anything

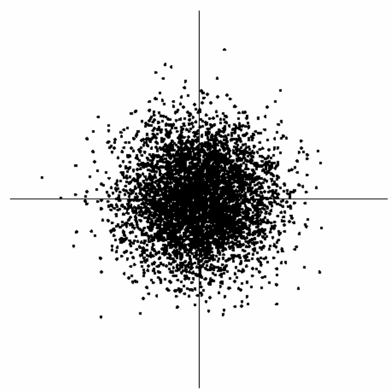
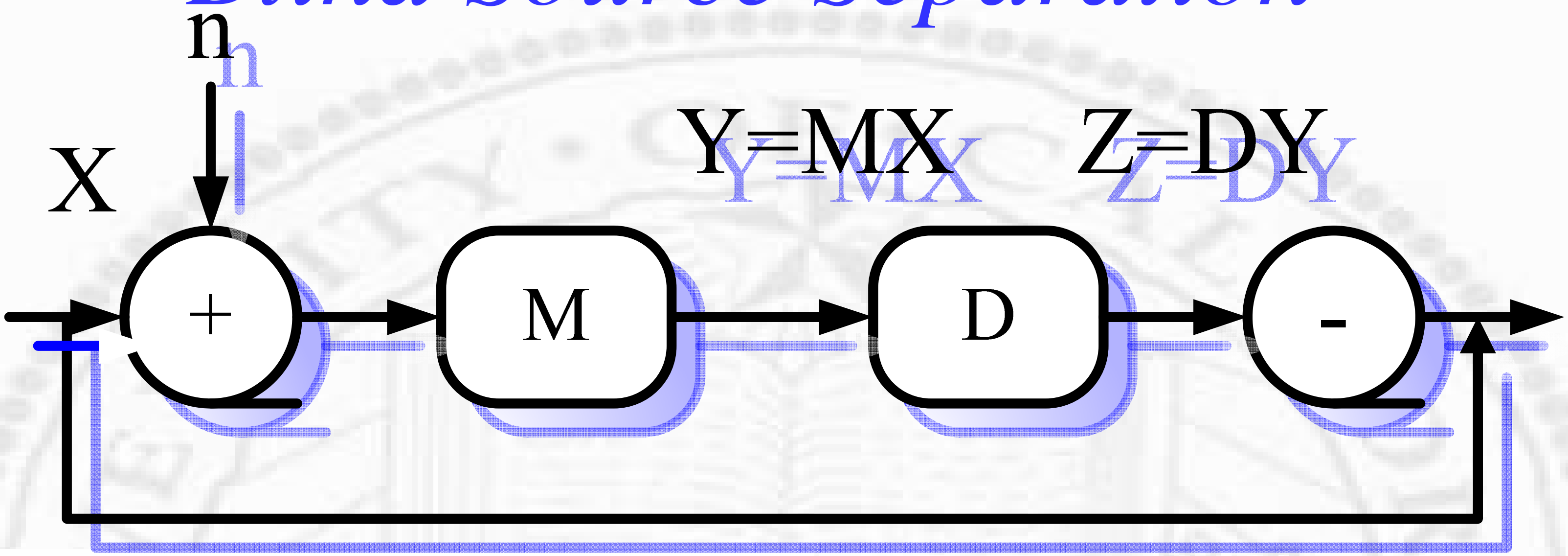
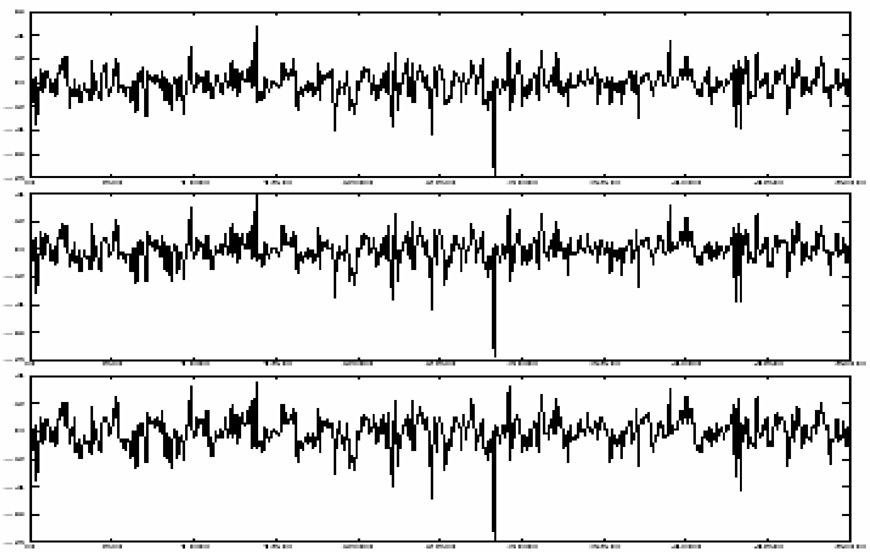
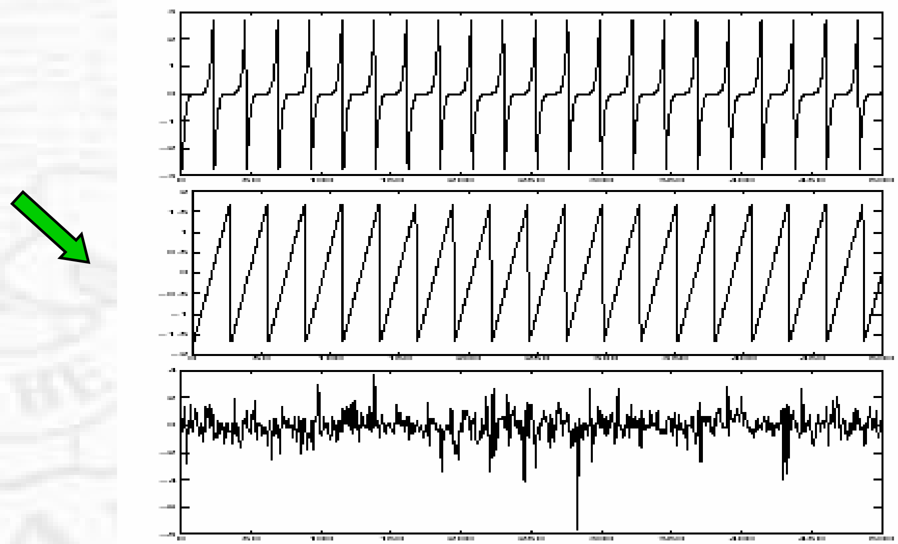


Figure 7: The multivariate distribution of two independent gaussian variables.

Blind Source Separation







Blind Source Separation

- Brain imaging
  - Different parts of brain emit signals that are mixed up in the sensors outside the bead
- Teleconferencing
  - Different speakers talk at the same time that are mixed up in the microphones
- Geology
  - Oil exploration with underground detonation and shock waves being registered at multiple sensors

Approaches

- Nonlinear de-correlation
  - The de-correlated components are uncorrelated and the transformed de-correlated components are uncorrelated
    - Minimum mutual information model
    - Maximum non-Gaussianity
- Maximum non-Gaussianity
  - Central limit theorem states more Gaussianity with successive mixture
    - Go above covariance matrix (kurtosis, a higher-order cumulant)

Mathematic Formulation

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jn}s_n, \text{ for all } j$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

$$\mathbf{s} = \mathbf{W}\mathbf{x}$$

- s_i: sources, x_j: mixtures
- **A**: mixture matrix
- **W**: de-mixing matrix
- Implication
  - Cannot determine the variance of sources
  - Cannot determine the ordering of source

A Simple Formulation

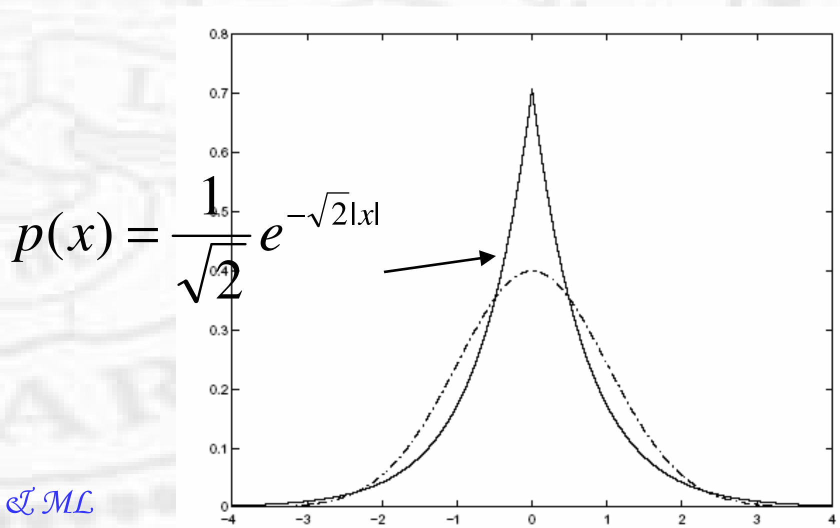
- Central Limit Theorem states that sum of independent random variables tends to Gaussian
- Non-Gaussianity is desired for each independent component

A Simple Formulation

- Gaussian variables have zero Kurtosis

$$kurt(x) = E(x^4) - 3(E(x^2))^2 = E(x^4) - 3 \quad if\ E(x^2) = 1$$

- Supergaussian: spiky pdf with heavy tails (e.g., Laplace distribution) $p(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|}$
- Subgaussian: flat pdf (e.g., uniform)
- Maximize magnitude of the Kurtosis



Math Framework: 2 variables 2 observations

For independent variables :

$$kurt(x_1 + x_2) = kurt(x_1) + kurt(x_2)$$

$$kurt(ax_1) = a^4 kurt(x_1)$$

$$y = \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{A}\mathbf{s} = \mathbf{z}^T \mathbf{s} = z_1 s_1 + z_2 s_2$$

$$\text{kurt}(y) = \text{kurt}(z_1 s_1) + \text{kurt}(z_2 s_2) = z_1^4 \text{kurt}(s_1) + z_2^4 \text{kurt}(s_2)$$

$$E\{y^2\} = z_1^2 + z_2^2 = 1$$

- All variables, s and y, are of unit variance
- Z is constrained to the unit circle
- Maximum kurtosis at two directions that lie in
  - z1=1 (-1), z2=0 or
  - z2=1 (-1) z1=0
- Through gradient search in **w**
- Drawback: noise sensitivity

Information

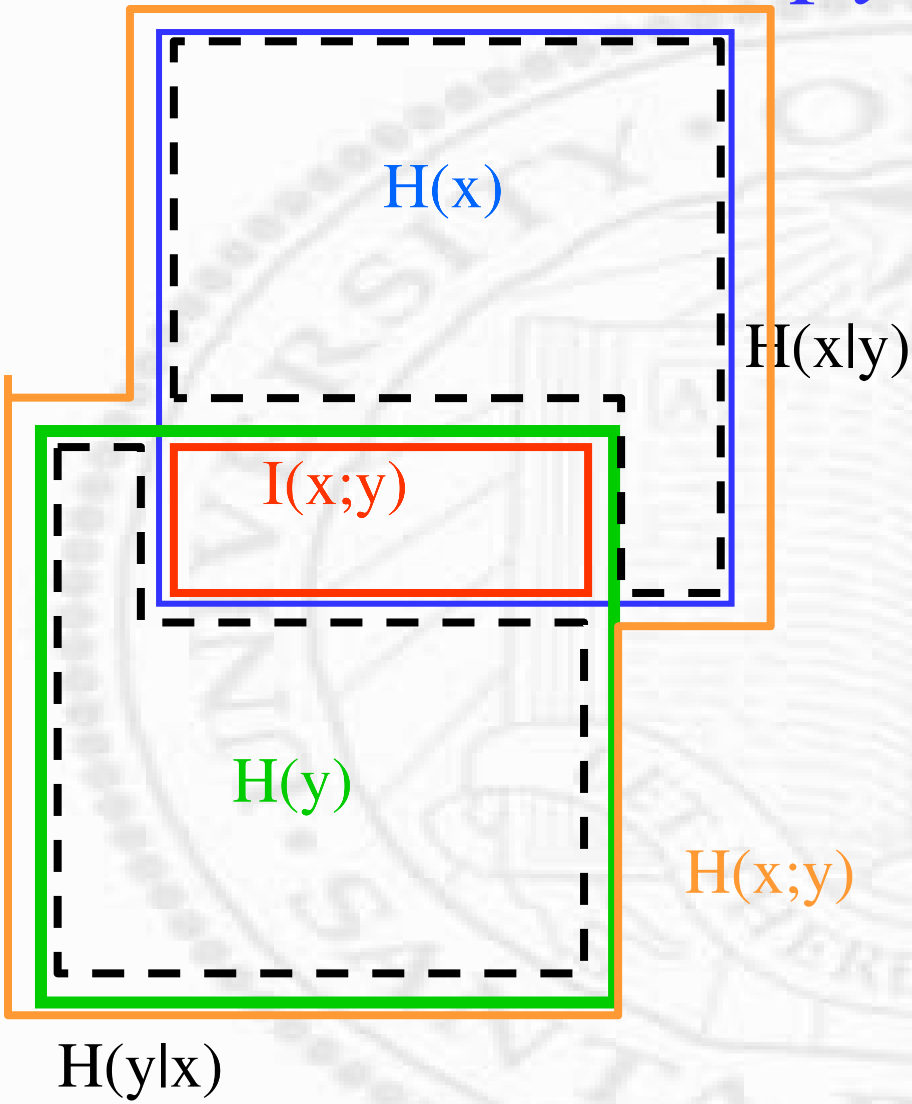
- Recall some important concepts
  - Random variable (**x**) $0 \le p_k = p(x = x_k) \le 1$
  - Probability distribution on a random variable
  - Amount of information, surprise, uncertainty

$$I(\mathbf{x} = \mathbf{x}_k) = \log(\frac{1}{p_k}) = -\log p_k$$

  - Entropy (weighted, average)

$$H(\mathbf{x}) = E(I(x_k)) = \sum_k p_k I(x_k) = -\sum_k p_k \log p_k$$

Entropy Basics



$$H[X,Y] \equiv -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2(\Pr(x,y))$$

$$H[X|Y] \equiv -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2 \Pr(x|y)$$

$$H[X,Y] = H[X] + H[Y|X]$$

$$H[X,Y] = H[Y] + H[X|Y]$$

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

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

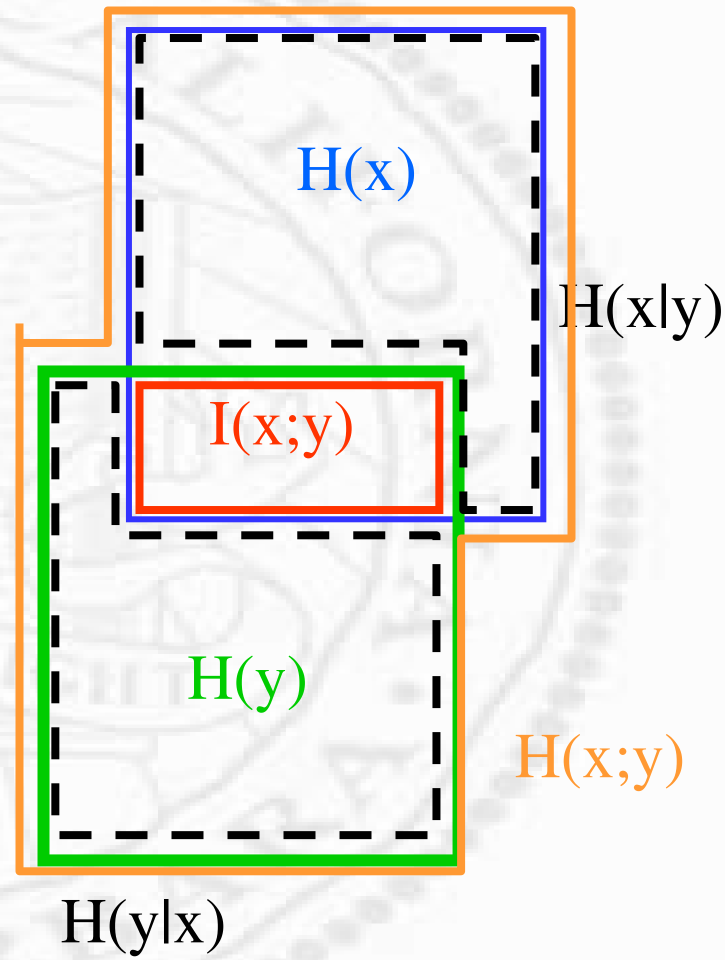
Mutual Information

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

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$$

$$\begin{aligned} I(X;Y) &= \sum_y p(y) \sum_x p(x|y) \log_2 \frac{p(x|y)}{p(x)} \\ &= \sum_y p(y)\, D_{\mathrm{KL}}(p(x|y) \| p(x)) \\ &= \mathbb{E}_Y\{D_{\mathrm{KL}}(p(x|y) \| p(x))\}. \end{aligned}$$



Kullback-Leibler divergence

$$D_{p\|q}(\mathbf{x}) = \sum_k p_k \log \frac{p_k}{q_k} = -\sum_k p_k \log q_k + \sum_k p_k \log p_k = H(p,q) - H(p)$$

- Information divergence, relative entropy
- Measure of difference between two distributions, but it is not a metric $D_{p\|q}(\mathbf{x}) \neq D_{q\|p}(\mathbf{x})$
- $D_{p\|q}$ is positive and is zero if and only if *p* and *q* have the same distribution
- Can be a useful measurement of independence, if
  - p is joint probability
  - q is marginal probability
- Then $D_{p\|q}$ is zero if and only if random variables are independent
- p = p(x,y) and q=p(x)p(y), the same as saying that x and y are independent

Intuition

- Independence implies product of marginal probabilities equals total probability

$$p(g_1(x_1), g_2(x_2), \cdots, g_n(x_n)) = p(g_1(x_1)) \cdots p(g_n(x_n))$$

$$p(x_1, x_2, \cdots, x_n) = p(x_1) \cdots p(x_n)$$

- The Kullback-Leibler divergence should be minimized

$$D_{p_{g(\mathbf{y})} \| p_{g(\tilde{\mathbf{y}})}} = \sum_k p_{g(\mathbf{y})=\mathbf{k}} \log \frac{p_{g(\mathbf{y})=\mathbf{k}}}{\prod_i p_{g(y_i)=k_i}}$$

$$D_{p_{\mathbf{y}} \| p_{\tilde{\mathbf{y}}}} = \sum_k p_{\mathbf{y}=\mathbf{k}} \log \frac{p_{\mathbf{y}=\mathbf{k}}}{\prod_i p_{y_i=k_i}}$$

Math Details

- A should minimize the mutual information between the new signal $H(Y_i)$ and the original signal $H(X)$

$$I(X) = \sum_i H(X_i) - H(X)$$

$$Y = AX$$

$$I(Y) = \sum_i H(Y_i) - H(X) - \log(\det A)$$

$$= \sum_i H(Y_i) - H(X)$$

Information Theoretic Approach

- Gaussian variable has the largest entropy among all variables of equal variance
- Negentropy (non-Gaussianality) *J* is to be maximized (X_{gauss} and *X* have the same variance)
  - $J(X) = H(X_{gauss}) - H(X)$
- Difficulty: computing *H* requires pdf
- Estimation:

$$J(x) \approx \frac{1}{12} E(x^3)^2 + \frac{1}{48} kurt(x)^2$$

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2$$

$1 \leq a_1 \leq 2$

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

Maximum Entropy Approach

$$p(\mathbf{x}(t)) = \prod_{i=1}^{d} p(x_i(t))$$

$$\mathbf{s}(t) = \mathbf{A}\mathbf{x}(t)$$

$$\mathbf{y}(t) = \mathbf{W}\mathbf{s}(t)$$

$$p_y(\mathbf{y}(t)) = \frac{p_s(\mathbf{s}(t))}{|\mathbf{J}|}$$