Shannon Information and Entropy in the Analysis of Independent Components and Clusters

Roman V. Belavkin

School of Science and Technology
Middlesex University, London NW4 4BT, UK

December 19, 2016

Dependencies in Data

Dependencies in data

- Consider the following records:

Case:	Age	Gender	Income (£ K)	Outcome (£ K)	Home owner	Credit score
1	21	0	2	1	0	3
2	18	1	1	2	0	1
3	50	1	6	2	1	5
4	23	0	3	1	1	4
5	40	1	3	2	0	2

- Each case is a vector $y \in \mathbb{R}^m$:

$$
\begin{aligned}
y^1 &= (21, 0, 2, 1, 0, 3)^T \\
y^2 &= (18, 1, 1, 2, 0, 1)^T \\
&\dots \\
y^n &= (23, 0, 3, 1, 1, 4)^T
\end{aligned}
$$

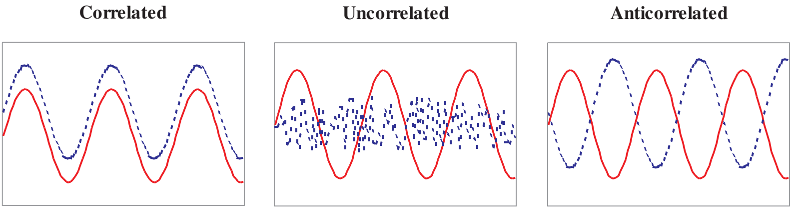
- The variables 'Age', 'Income', 'Outcome' define a basis in $\mathbb{R}^m$, and we are interested in dependencies between the variables.

Correlation

- Correlation is the measure of linear dependency:

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var\{x\}Var\{y\}}}$$

- If $x = y$, then $Corr(x, y) = 1$ (for $Cov(x, x) = Var\{x\}$)



$Corr(x, y) = 1$ $\quad$ $Corr(x, y) = 0$ $\quad$ $Corr(x, y) = -1$

Correlation matrix

	Age	Gender	Income	Outcome	H. owner	C. score
Age	1,0	0,6	0,9	0,6	0,4	0,5
Gender	0,6	1,0	0,2	1,0	-0,2	-0,3
Income	0,9	0,2	1,0	0,2	0,7	0,9
Outcome	0,6	1,0	0,2	1,0	-0,2	-0,3
H. owner	0,4	-0,2	0,7	-0,2	1,0	0,9
C. score	0,5	-0,3	0,9	-0,3	0,9	1,0

Principle component analysis

- PCA is a linear transformation of data $y \mapsto Ky = x$:

$$Ky = \begin{pmatrix} k_{11} & \dots & k_{1m} \\ \vdots & \ddots & \vdots \\ k_{m1} & \dots & k_{mm} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = x$$

- Such that the transformed vectors $x = (x_1, \dots, x_m)$ have uncorrelated coordinates:

$$Corr(x_i, x_j) = 0 \quad \text{for all} \quad i \neq j$$

- Often most of the variance in the data is accounted by variance in only a few ($k < m$) components (the principal components).

Correlation $\not\equiv$ dependency

- Let $x \in \mathbb{R}$ and y be defined as:

$$y = \sin(x)$$

- Thus, y depends on x by functionally, but

$$Corr(x, y) = 0$$

- To see this, recall that correlation represents an average linear trend between y and x.
- Generally

$$\begin{aligned} x, y \text{ are independent} \quad &\Rightarrow \quad Corr(x, y) = 0 \\ x, y \text{ are independent} \quad &\not\Leftarrow \quad Corr(x, y) = 0 \end{aligned}$$

Independence

- Recall that x and y are independent if and only if the conditional probability $P(y \mid x)$ equals to $P(y)$ (marginal):

$$P(y \mid x) = P(y) \quad \text{or} \quad J(x, y) = Q(x)P(y)$$

- Dependency is measured by mutual information:

$$I(x, y) := \mathbb{E}_J \left\{ \ln \frac{P(y \mid x)}{P(y)} \right\} = \sum_{x,y} \left[\ln \frac{P(y \mid x)}{P(y)} \right] J(x, y) \geq 0$$

- For dependency in $y = (y_1, \ldots, y_m)$ we can consider the divergence:

$$I(y_1, \ldots, y_m) = \sum_{y_1, \ldots, y_m} \left[\ln \frac{J(y_1, \ldots, y_m)}{P(y_1) \otimes \cdots \otimes P(y_m)} \right] J(y_1, \ldots, y_m) \geq 0$$

Information as distance

- Kullback-Leibler divergence of Q from P in $\mathcal{P}$:

$$D_{KL}[P,Q] := \mathbb{E}_P\{\ln P - \ln Q\} = \sum_{\Omega} [\ln P(\omega) - \ln Q(\omega)]\ P(\omega)$$

- Surprise associated with observation of event $e \in \Omega$:

$$D_{KL}[\delta_e, Q] = \sum_{\Omega} \left[\ln \delta_e(\omega) - \ln Q(\omega)\right]\ \delta_e(\omega) = -\ln Q(e)$$

- Entropy is expected surprise

$$H[Q] := \mathbb{E}_Q\{-\ln Q\} = -\sum [\ln Q]\, Q$$

- Shannon (1948) information is divergence of product of marginals $Q \otimes P$ from joint measure J:

$$D_{KL}[J, Q \otimes P] = \mathbb{E}_J\{\ln J - \ln Q \otimes P\} =: I(x,y)$$

Shannon information and entropy

- Shannon (1948) mutual information between x and y:

$$
\begin{aligned}
I(x, y) &= \sum_{X \times Y} \left[\ln \frac{J(x, y)}{Q(x)\, P(y)} \right] J(x, y) \\
&= \sum_{Y} P(y) \sum_{X} \left[\ln \frac{Q(x \mid y)}{Q(x)} \right] Q(x \mid y) \\
&= H[Q(x)] - H[Q(x \mid y)] \\
&= H[P(y)] - H[P(y \mid x)] \\
&= H[Q(x)] + H[P(y)] - H[J(x, y)]
\end{aligned}
$$

- Shannon information of x is:

$$I(x, x) = H[Q]$$

- If x has elementary distribution $\delta_\omega(E)$, then:

$$I(x, x) = H[\delta] = 0$$

Blind source separation

- ICA belongs to a class of techniques for *blind source separation*
- The data $y \in \mathbb{R}^m$ that we observe is the result of some unknown transformation f of some unobserved source signals $x \in \mathbb{R}^n$:

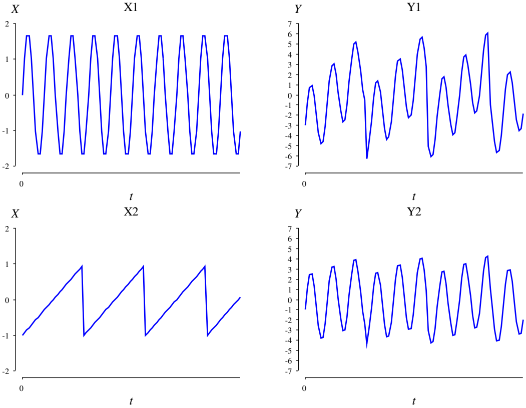
$$y = f(x)$$

- The goal of BSS is to find the inverse transformation f^{-1} (and hence the sources $x = f^{-1}(y)$) only based on the observed data.
- BSS is possible under some assumptions, such as if f is a linear transformation of independent sources:

$$y = Mx\,, \quad W \approx M^{-1}\,, \quad x \approx Wy$$

Example: The cocktail party problem

- The sources $x = (x_1, \dots, x_m)$ are m people at a party, whose voices are recorded by $n \geq m$ microphones.
- The data $y = (y_1, \dots, y_n)$ are n recordings of mixed signals.



Independent component analysis

- ICA is a linear transformation of data $y \mapsto Wy = x$:

$$Wy = \begin{pmatrix} w_{11} & \dots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \dots & w_{mm} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = x$$

- Such that the transformed vectors $x = (x_1, \dots, x_m)$ have independent coordinates:

$$J(x_1, \dots, x_m) = P(x_1) \otimes \cdots \otimes P(x_m) \quad \text{or} \quad I(x_1, \dots, x_m) = 0$$

- This can be achieved by iterative algorithms that estimate matrix W minimizing $I(Wy)$.

FastICA algorithm

- Recall the Central Limit Theorem, according to which the sum $x_1 + \cdots + x_n$ of n independent random variables with essentially bounded variances converges (in distribution) to a Gaussian random variable.
- Thus, the observed data $y_i = w_{i1}x_1 + \cdots + w_{im}x_m$ is generally 'more Gaussian' than the independent sources x_j.
- The non-Gaussianity is measured by neg-entropy, which is approximated by

$$I(y_i) = |\mathbb{E}\{G(y_i)\} - \mathbb{E}\{G(v)\}|^2$$

  where v is normal $N(0,1)$ and G are special functions (e.g. $G(u) = (1/\alpha)\log\cosh(\alpha u)$ or $G(u) = -\exp(u^2/2)$)
- The FastICA algorithm (Hyvärinen & Oja, 1997) iteratively finds $W \approx M^{-1}$ maximizing $I(y_i)$ for Wy

Direct entropy minimization algorithms

- The divergence $I(Wy) = I(x_1, \dots, x_n)$ in terms of entropies:

$$I(Wy) = \sum_{x_1,\dots,x_n} \left[\ln \frac{J(x_1,\dots,x_n)}{P(x_1)\otimes\cdots\otimes P(x_n)}\right] J(x_1,\dots,x_n)$$

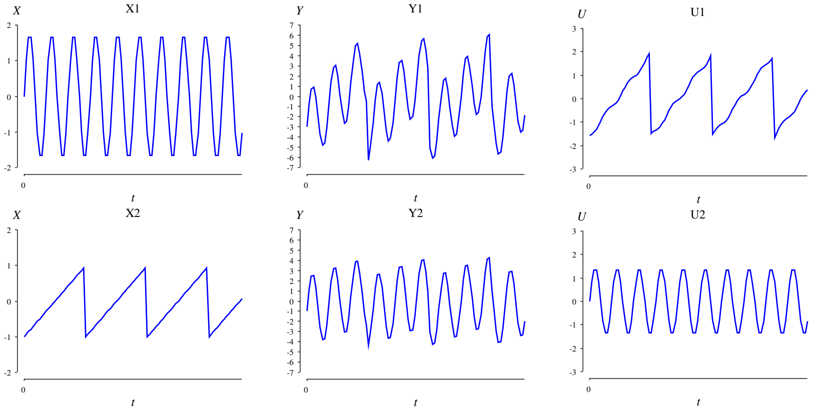
$$= -\left(\sum_{x_1}[\ln P(x_1)]\,P(x_1) + \cdots + \sum_{x_n}[\ln P(x_n)]\,P(x_n)\right)$$

$$+ \sum_{x_1,\dots,x_n}[\ln J(x_1,\dots,x_n)]\,J(x_1\ \dots,x_n) = \sum_{i=1}^{n} H[P(x_i)] - H[J(Wy)]$$

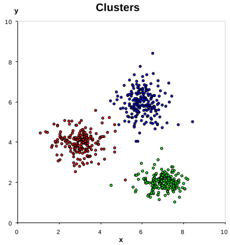
- If $Wy = x$ is injective, then $H[J(y)] = H[J(Wy)] = H[J(x)]$, so that

$$\min_{Wy=x} I(Wy) \qquad \Longleftrightarrow \qquad \max_{Wy=x} \sum_{i=1}^{m} H[P(x_i)]$$

- RADICAL algorithm does this using Jacobi rotations (Learned-Miller & Fisher, 2003) and using ordered statistics (Vasicek, 1976) to estimate $H[P(x_i)]$.


X1
X
t
Y1
Y
t
U1
U
t
X2
X
t
Y2
Y
t
U2
U
t

Clustering



- **Clustering** is a partition $X = X_1 \cup \cdots \cup X_k$ of data.
- It is a mapping $f : X \to Y$ to a set Y of labels (codes):

$$x \mapsto f(x) = y$$

- The groups can be based on similarity.

Example (k-means)

Y is the set of k points in (X, d), and $f : X \to Y$ solves:

$$\min_{f(x)=y} \sum_{i=1}^{k} \sum_{x \inf^{-1}(y_i)} d(x, y_i)$$

The new $y_i \in Y$ are set to be the centroids $y_i^{t+1} = \mathbb{E}\{x \inf^{-1}(y_i^t)\}$.

Clustering as source coding

- $f : X \to Y$ is an encoding, where each y_i must have as much information about $x \in X_i = f^{-1}(y_i)$ as possible.
- Trivial solution is to use an injective (or uniquely-decodeable) code:

$$f(x_i) = f(x_j) \qquad \Rightarrow \qquad x_i = x_j$$

- Usually, we want some compression $k = |Y| \ll |X|$ (non-injective f).
- and preserving as much information as possible:

$$\max_{f(x)=y} I(x, y)$$

Conditional entropy minimization clustering

- for $f(x) = y$

$$P(y \mid x) = \delta_{f(x)}(y) = \begin{cases} 1 & \text{if } y = f(x) \\ 0 & \text{otherwise} \end{cases}$$

- Conditional entropy $H[P(y \mid x)] = 0$

$$I(x, y) = H[P(y)] \leq \ln |Y|$$

- Maximize $H[P(y)] \leq H[P(x)]$:

$$k = |Y| \leq e^{H[P(x)]}$$

- for $x \inf^{-1}(y)$

$$Q(x \mid y) = \frac{Q(x)}{\sum_{x \inf^{-1}(y)} Q(x)}$$

- Conditional entropy $H[Q(x \mid y)] \geq 0$

$$I(x, y) = H[Q(x)] - H[Q(x \mid y)]$$

- Minimize $H[Q(x \mid y)]$:

$$H[Q(x \mid y)] = \sum^{k} H[Q(x \overset{-1}{\inf}(y_i))]$$

Detection of HTTP-GET attack

Entropy-based clustering of user online behaviour (Chwalinski, Belavkin, & Cheng, 2013).

Value of information and optimal solutions

- Linear programming problem to find optimal $\hat{P}(y \mid x) = \frac{\hat{J}(x,y)}{Q(x)}$:

$$\text{minimize} \quad \mathbb{E}_J\{d(x,y)\} \quad \text{subject to} \quad I(x,y) \leq \lambda$$

- The inverse convex programming problem:

$$\text{minimize} \quad I(x,y) \quad \text{subject to} \quad \mathbb{E}_J\{d(x,y)\} \leq \upsilon$$

- Optimal solution for $d(x+a, y+a) = d(x,y)$ (Stratonovich, 1975):

$$\hat{Q}(x \mid y) = \frac{e^{-\beta d(x,y)}}{\sum_X e^{-\beta d(x,y)}}, \qquad \beta^{-1} = -\frac{d}{d\lambda}\mathbb{E}_{\hat{J}}\{d\}(\lambda)$$

- Optimal transformation $x \mapsto y$ given by $\hat{P}(y \mid x)$ is randomized (Belavkin, 2013).

Geometric value of information

- $\mathbb{E}_p\{u\} = \langle u, p \rangle$ expected utility
- $F[p,q]$ information divergence
- Value of information λ:

$$v_u(\lambda) := \sup\{\langle u, p \rangle : F[p,q] \leq \lambda\}$$

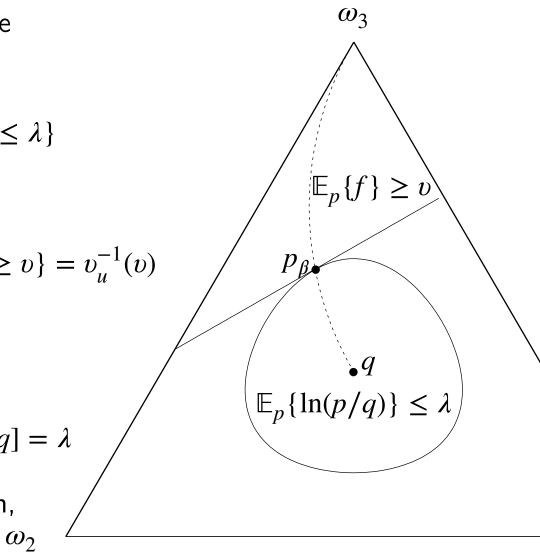
- Information of value v:

$$\lambda_u(v) := \inf\{F[p,q] : \langle u, p \rangle \geq v\} = v_u^{-1}(v)$$

- Optimal solutions:

$$p(\beta) \in \partial F^*[\beta u, q], \quad F[p(\beta), q] = \lambda$$

- (Stratonovich, 1965; Belavkin, 2013)



Dependencies in Data

Shannon's Information and Entropy

Independent Component Analysis

Clustering

Value of Information

Belavkin, R. V. (2013). Optimal measures and Markov transition kernels. *Journal of Global Optimization*, *55*, 387–416.

Chwalinski, P., Belavkin, R., & Cheng, X. (2013). Detection of HTTP-GET attack with clustering and information theoretic measurements. In *Foundations and practice of security* (Vol. 7743, pp. 45–61). Springer Berlin Heidelberg.

Hyvärinen, A., & Oja, E. (1997). A fast fixed point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492.

Learned-Miller, E., & Fisher, J. (2003). ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, *4*, 1271–1295.

Shannon, C. E. (1948, July and October). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423 and 623–656.

Stratonovich, R. L. (1965). On value of information. *Izvestiya of USSR Academy of Sciences, Technical Cybernetics*, *5*, 3–12. (In Russian)

Stratonovich, R. L. (1975). *Information theory*. Moscow, USSR: Sovetskoe Radio. (In Russian)

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, *38*(1), 54–59.