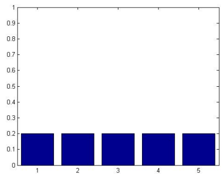
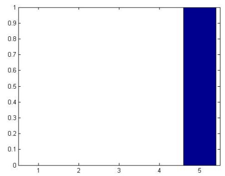
Entropy and related information theoretical measures: Examples on application in data analysis

Allan Aasbjerg Nielsen
Technical University of Denmark
DTU Compute – Department of Applied Mathematics and Computer Science
alan@dtu.dk

IACG Christmas WS 17 Dec 2013

Entropy



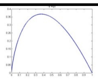


max entropy $= \ln(n) = \ln(5) = 1.6094$ min entropy = 0

Entropy

- Discrete stochastic variable X with probability density function (pdf) $p(X = x_i)$, $i = 1, \dots, n; n$ is number of possible outcomes (or bins)
- Measure of ***information*** (or ***surprise***) $h(X = x_i)$: x_i very probable, i.e., $p(X = x_i)$ is high then $h(X = x_i)$ should be low, and *v.v.*
- Realizations x_i and y_j of independent X and Y, i.e., the two-dimensional pdf $p(X = x_i, Y = y_j)$ equals the product of the one-dimensional marginal pdfs $p(X = x_i)p(Y = y_j)$, we would like the joint information content to equal the sum of the marginal information contents, i.e., $h(X = x_i, Y = y_j) = h(X = x_i) + h(Y = y_j)$
- Feasible with $h(X = x_i) = -\ln(p(X = x_i))$
- The expectation $H(X)$ of the information content is termed the (Shannon) ***entropy***: $H(X) = -\sum_{i=1}^{n} p(X = x_i)\ln(p(X = x_i)) = -\mathrm{E}\{\ln(p)\}$

Entropy

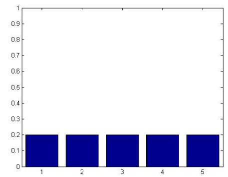


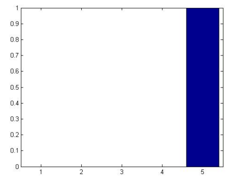
- $\lim_{p \to 0+} p\ln(p) = 0$
- Unit: nat, nit or nepit (natural log); bit (log base 2); ban or dit (log base 10)
- For continuous stochastic variable $H(X) = -\int p(x)\ln(p(x))\,dx$
  - Differential entropy; since $p(x)$ can be >1, $H(X)$ can be negative (or infinite)
  - Gaussian pdf has max entropy for continuous distributions with finite variance, $H(X) = \{1 + \ln(2\pi\sigma^2)\}/2$; $H(\boldsymbol{X}) = \ln(|(2\pi e)\boldsymbol{\Sigma}|)/2$
- Discrete X: be careful if binning is applied
- Empirical entropy $\widehat{H}(X) = -\sum_{i=1}^{N} \ln(p(X = x_i))/N$; N is number of obs
- Rényi entropy: $H_\alpha(X) = \frac{1}{1-\alpha}\ln(\sum_{i=1}^{n} p(X = x_i)^\alpha)$, $\alpha > 0$ and $\alpha \neq 1$
  - $\alpha \to 1$: Shannon entropy
  - $\alpha = 2$: Collision or ***Rényi entropy***: $H_2(X) = -\ln(\sum_{i=1}^{n} p(X = x_i)^2) = -\ln(\mathrm{E}\{p\})$

Entropy

- ***Shannon entropy:*** $-\mathbf{E}\{\ln(p)\}$
- ***Rényi entropy:*** $-\ln(\mathbf{E}\{p\})$ $\mathbf{E}\{p\}$ ***is energy***
- ***Entropy is the average amount of information (or surprise) obtained from obs***
- ***Entropy is a measure of order***
  - ***Physical system: heat source/sink, ability to do work, low entropy***
    Source/sink same temperature: no work done, max entropy
    Histogram of temperature of source/sink
  - ***A bit like my (low entropy) office: a lot of potential for work***

Entropy





max entropy $= \ln(n) = \ln(5) = 1.6094$ min entropy = 0

Entropy

- Claude Elwood Shannon (1916-2001), A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948: ***application areas communication, compression*** – "father of information theory"
- Alfréd Rényi (1921-1970), contributions in combinatory, graph theory, number theory but mostly in probability theory
- John von Neumann (1903-1957), major contributions in mathematics, physics, economics, computer science, and statistics (source http://en.wikipedia.org/wiki/John_von_Neumann): use the term entropy not only because of the link to physics, but also because ***"nobody knows what entropy really is, so in any discussion you will always have an advantage"***, (Christopher Michael Bishop, 2006)

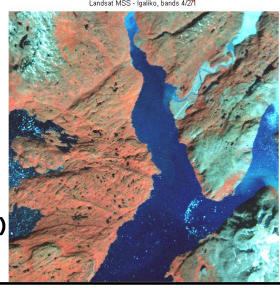
Entropy

- Early use in maximum entropy (image) reconstruction (B. Roy Frieden, 1972)
  - Entropy as regularizer in inverse problems to
  - suppress low-intensity ripple
  - sharpen point sources
  - in e.g. astronomical data
- Equality of distributions of saleries/income
  - Gini index – first order series expansion of $\ln(x)$ in Shannon entropy around $x = 1$, $\ln(x) \cong x - 1$: Gini $= \sum_{i=1}^{n} p(X = x_i)(1 - p(X = x_i)) = \mathrm{E}\{1 - p\}$
- None of this described further here, on to simpler things

Entropy of Posterior Probability

Source: http://www.denstoredanske.dk

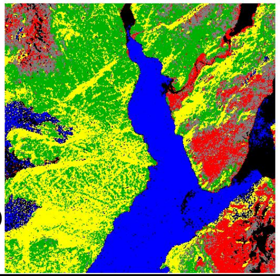
Landsat MSS - Igaliko, bands 4/2/1



Igaliko
Brattahlið (Qassiarsuk)
Landsat MSS

Entropy of Posterior Probability

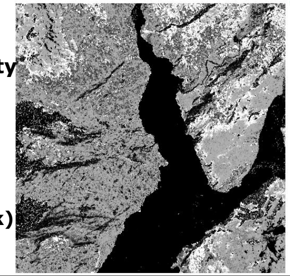
ML classification with reject class



Igaliko
Brattahlið (Qassiarsuk)
Landsat MSS

Entropy of Posterior Probability

Entropy of posterior probability to characterize certainty of classification

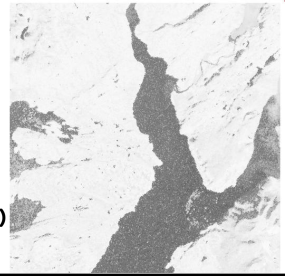


Igaliko
Brattahlið (Qassiarsuk)
Landsat MSS

Spectral Information Measure

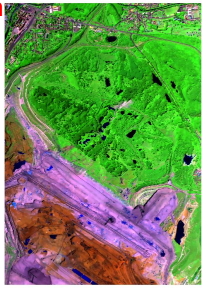
Entropy of DNs to characterize spectral information content

Chen-I. Chang, 2000



Igaliko
Brattahlið (Qassiarsuk)
Landsat MSS

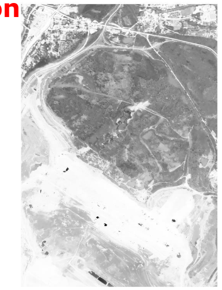
Spectral Information Measure



Sokolov
HyMap, 110 bands

Spectral Information Measure

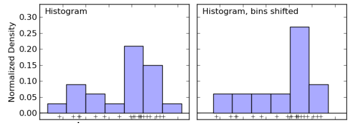
Entropy of DNs to characterize spectral information content



Sokolov
HyMap, 110 bands

Density Estimation

- Discrete X: be careful if binning is applied
  - The estimated histogram is not smooth and it depends on the end points of bins and the width of bins
  - Using kernel density estimators where we center a kernel on each observation, we may obtain smoother histograms that do not depend on bin end points
- Kernel density estimat value t is $\sum_{i=1}^{N} \varphi(t - x$ $\varphi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{t}{2}$
  - Issue: determine banc
- This parameterization

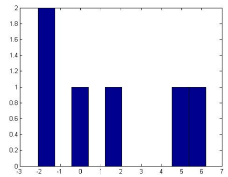


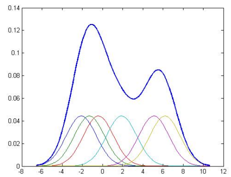
Density Estimation

- Discrete X: be careful if binning is applied
  - The estimated histogram is not smooth and it depends on the end points of bins and the width of bins
  - Using kernel density estimators where we center a kernel on each observation, we may obtain smoother histograms that do not depend on bin end points
- Kernel density estimator (Parzen window estimator) for the pdf of X at value t is $\sum_{i=1}^{N} \varphi(t - x_i)/N$; often with Gaussian kernel
  $\varphi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{t^2}{2\sigma^2})$ (Bernard W. Silverman, 1986)
  - Issue: determine bandwidth σ
- This parameterization facilitates optimization

Density Estimation

$x = [-2.1 \quad -1.3 \quad -0.4 \quad 1.9 \quad 5.1 \quad 6.2\,]$





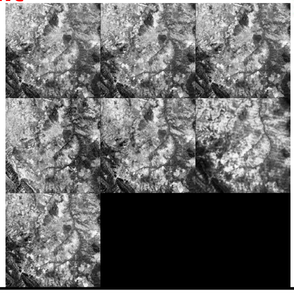
Source: http://en.wikipedia.org/wiki/Kernel_density_estimation

Projection Pursuit

- PCA (Harold Hotelling, 1933): linear combinations $\boldsymbol{a}^T\boldsymbol{X}$ of m (de-meaned) variables in $\boldsymbol{X}$ with max variance
  - Higher order components orthogonal to lower order components
- PP (Jerome H. Friedman and John Tukey, 1974): linear combinations $\boldsymbol{a}^T\boldsymbol{X}$ of m (de-meaned) variables in $\boldsymbol{X}$ with max deviation from uniform/Gaussian distribution after mapping with essentially Φ (standard normal cdf) measured by integral-squared distance or with min entropy
  - Higher order components: replace structure in projections with Gaussian noise and go again, i.e., no orthogonality, more than m projections
- Quite similar to ICA (Pierre Comon, 1994)

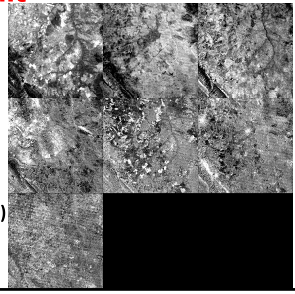
Projection Pursuit

Landsat TM data
from Central Spain
Bands 1-7



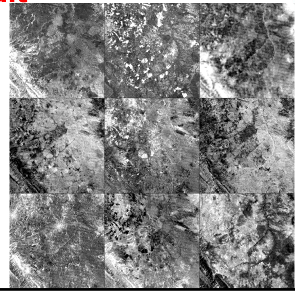
Projection Pursuit

Landsat TM data
from Central Spain
Principal components (corr.)



Projection Pursuit

Landsat TM data
from Central Spain
PP, 9 projections
(Kristian Windfeld)



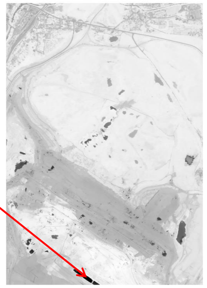
Relative Entropy/KL divergence

- ***Relative entropy*** aka ***Kullback-Leibler divergence*** between two pdfs $p(X = x_i)$ and $q(X = x_i)$ defined on the same set of outcomes (or bins): the expectation of the logarithmic difference between p and q
  $D_{KL}(p,q) = \sum_{i=1}^{n} p(X = x_i) \ln \frac{p(X=x_i)}{q(X=x_i)}$
- Measure of proximity, satisfies Gibbs' inequality $D_{KL}(p,q) \geq 0, = 0$ for $p(X = x_i) = q(X = x_i)$ only; not symmetric in p and q (not a metric)
- Symmetrized measure: $D_{KL}(p,q) + D_{KL}(q,p)$
- ***Rényi divergence***
  $D_\alpha(p,q) = \frac{1}{\alpha-1} \ln\left(\sum_{i=1}^{n} \left(\frac{p(X=x_i)^\alpha}{q(X=x_i)^{\alpha-1}}\right)\right) = \frac{1}{\alpha-1} \ln(\sum_{i=1}^{n} p(X = x_i)^\alpha q(X = x_i)^{1-\alpha})$

Spectral Angle

Inverse cos of dot pr to characterize spectral information differences, target in bottom lake

Sokolov
HyMap, 110 bands

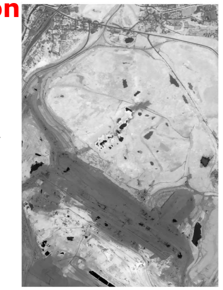


Roberta H. Yuhas, Alexander F.H. Goetz, and Joe W. Boardman, 1992

Spectral Information Divergence

Symm KL div of pdfs to characterize spectral information differences, target in bottom lake

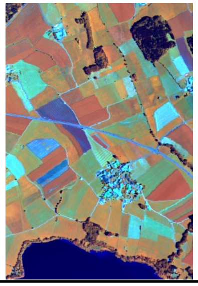
Sokolov
HyMap, 110 bands

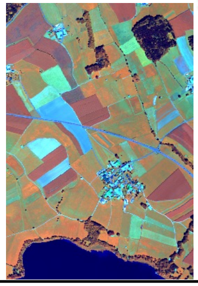


Chen-I. Chang, 2000

Change Detection: SAM, KL

Lake Waging-Taching,
HyMap, 126 bands
30 Jun 2003 8:43 UTC
4 Aug 2003 10:23 UTC





Spectral Angle

Inverse cos of dot pr to characterize spectral change

Lake Waging-Taching
HyMap, 126 bands

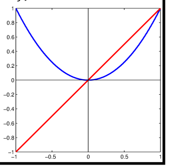
Spectral Information Divergence

Symm KL div of pdfs to characterize spectral change

Lake Waging-Taching
HyMap, 126 bands

Mutual Information

- ***Kullback-Leibler divergence***
  $D_{KL}(p,q) = \sum_{i=1}^{n} p(X = x_i) \ln \frac{p(X=x_i)}{q(X=x_i)}$
- ***Mutual information***: measure of extent to which X and Y are independent, KL divergence between $p(X = x_i, Y = y_j)$ and $p(X = x_i)p(Y = y_j)$
  $I(X,Y) = \sum_{i,j} p(X = x_i, Y = y_j) \ln \frac{p(X=x_i, Y=y_j)}{p(X=x_i)p(Y=y_j)}$
- $I(X,Y)$ symmetric in X and Y (a metric)
- $I(X,Y) = H(X) + H(Y) - H(X,Y)$
- A normed version as supplement to correlation

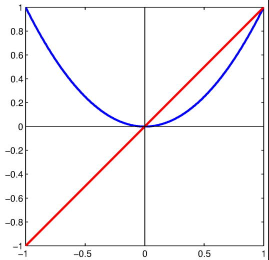


Mutual Information

- CCA: linear combinations $\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{b}^T\boldsymbol{Y}$ of m and k (de-meaned) variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ with max correlation (Harold Hotelling, 1936)
  - Higher order components orthogonal to lower order components
- ***Replace correlation in CCA with mutual information*** (Xiangrong Yin, 2004; Karasuyama & Sugiyama, 2012; Vestergaard & Nielsen: CIA, 2015)
  - Higher order components: min MI with lower order components; no orthogonality; more than $\max(m,k)$ components
  - CIA handles large datasets
- Examples
- CIA paper

Toy example

- $\boldsymbol{X} = [x_1 \;\; x_2]$, $\boldsymbol{Y} = [y_1 \;\; y_2]$
- x_1 sampled equidistantly on $[-1 \;\; 1]$
- $y_1 = x_1^2 + \varepsilon_1$
- $x_2 = \varepsilon_2$
- $y_2 = \varepsilon_3$
- $\varepsilon_i \sim N(0, 0.1^2)$
- max $I(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{Y})$



Toy example