JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Renyi's entropy

History: Alfred Renyi was looking for the most general definition of information measures that would preserve the additivity for independent events and was compatible with the axioms of probability.

He started with Cauchy's functional equation: If p and q are independent than I(pq)=I(p)+I(q).

Apart from a normalizing constant this is compatible with Hartley's information content I(p)=-log p. If we assume that the events $X=\{x_1,...x_N\}$ have different probabilities $\{p_1,...p_N\}$, and each delivers I_k bits of information, then the total amount of information for the set is

$$I(P)=\sum_{k=1}^{N} p_k I_k$$

This can be recognized as Shannon's entropy. But he reasoned that there is an implicit assumption used in this equation: we use the linear average, which is not the only one that can be used.

In the general theory of means, for any function g(x) with inverse g^{-1}, the mean can be computed as

$$g^{-1}(\sum_{k=1}^{N} p_k g(x_k))$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Applying this definition to the I(P) we get

$$I(P) = g^{-1}(\sum_{k=1}^{N} p_k g(I_k))$$

When the postulate of additivity for independent events is applied we get just two possible g(x):

$$g(x)=cx$$

$$g(x)=c^{-2(1-\alpha)x}$$

The first form gives Shannon information and the second gives

$$I_\alpha(P) = \frac{1}{1-\alpha}\log(\sum_{k=1}^{N} p_k^\alpha)$$

for non negative α different from 1. This gives a parametric family of information measures that are called today Renyi's entropies.

It can be shown that Shannon is a special case when $\alpha \to 1$

$$\lim_{\alpha\to 1} H_\alpha(X) = \lim_{\alpha\to 1}\frac{1}{1-\alpha}\log\sum_{k=1}^{N} p_k^\alpha = \frac{\lim_{\alpha\to 1}\sum_{k=1}^{N} \log p_k \cdot p_k^\alpha \Big/ \sum_{k=1}^{N} p_k^{\alpha-1}}{\lim_{\alpha\to 1} -1} = H_S(X)$$

So Renyi's entropies contain Shannon as a special case.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

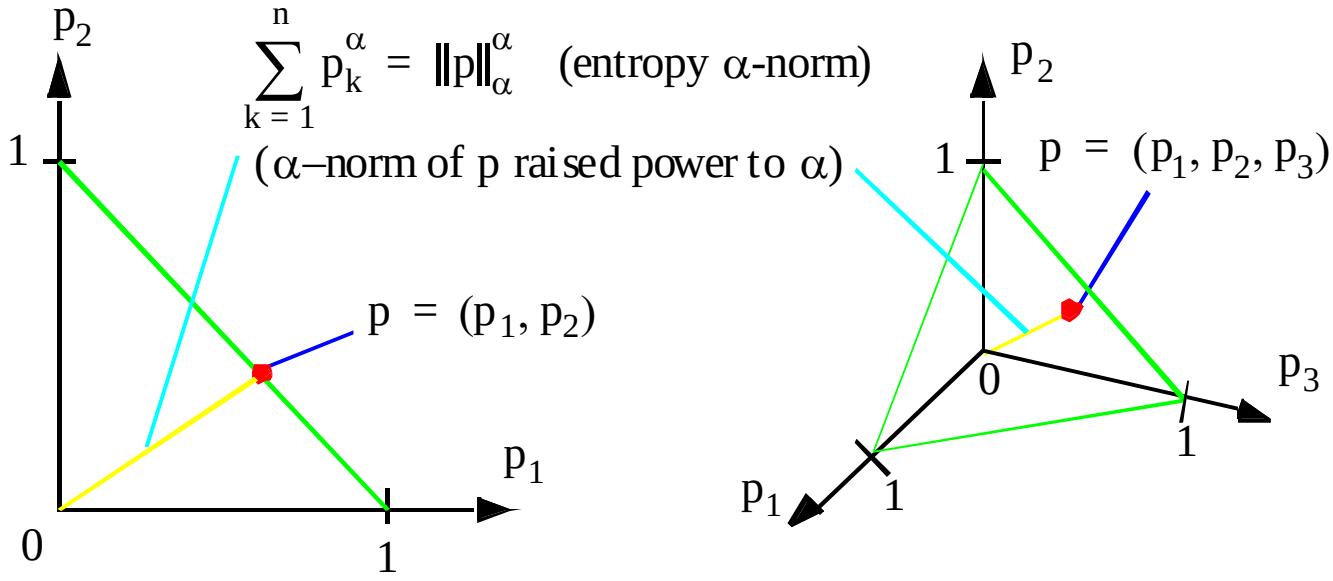
Meaning of α

If we compare the two definitions we see that instead of weighting the log p_k by the probabilities, here the log is external to the sum and its argument is the α power of the PMF, i.e.

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(V_\alpha(X)\right) = -\log\left(\sqrt[\alpha-1]{V_\alpha(X)}\right) = -\log\left(\sqrt[\alpha-1]{E(V_{\alpha-1}(X))}\right) \qquad V_\alpha(X) = \sum_k p_k^\alpha$$

Note that Vα(X) is the argument of the α norm of the PMF.

A geometric view will help here. All the PMFs of N r.v. exist on what is called the simplex, in an N dimensional space.



JOSE C. PRINCIPE

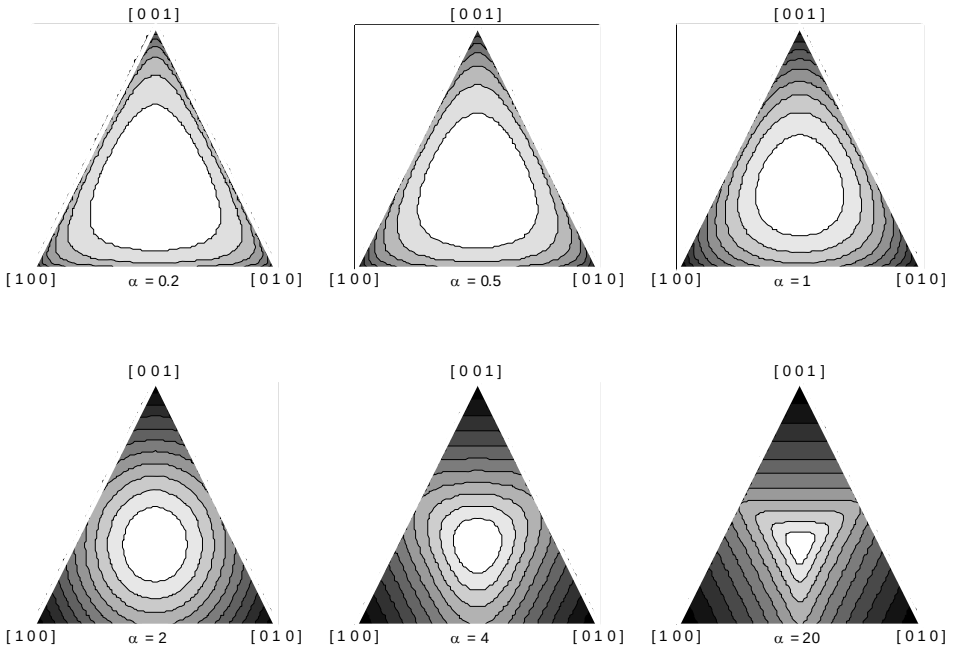
U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

The norm measures exactly the distance of the PMF to the origin, and α is the designated l-norm (Euclidean norm is l=2). Changing α changes the metric distance in the simplex.



From another perspective, the α root of $V_\alpha(X)$ is the α-norm of the PMF. So we conclude that Renyi's entropy is more flexible than Shannon and includes Shannon as a special case.

We will be using extensively α=2.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Properties of Renyi's entropies

(a) $H_\alpha(X)$ is nonnegative: $H_\alpha(X) \geq 0$;

(b) $H_\alpha(X)$ is decisive: $H_\alpha(0, 1) = H_\alpha(1, 0)$;

(c) For $\alpha \leq 1$ Renyi's entropy is concave. For $\alpha > 1$ Renyi's entropy in not pure convex nor pure concave ; It looses concavity for $\alpha > \alpha^* > 1$ where α^* depends on N and obeys the relation $\alpha^* \leq 1 + \ln(4)/\ln(N-1)$.

(d) Since $\frac{\alpha-1}{\alpha} H_\alpha(X) \leq \frac{\beta-1}{\beta} H_\beta(X)$ *for* $\alpha \leq \beta$, $(\alpha\text{-}1)H_\alpha(X)$ is a concave function of X.

(e) $H_\alpha(X)$ is bounded, continuous and non increasing function of α;

(f) Renyi's entropies for different α are correlated.

(g) The following is a simple but not very sharp bound on Shannon entropy ($H_S(X)$) of any probability mass function

$$H_2(X) \leq H_S(X) \leq \ln N + 1/N - \exp(-H_2(X))$$

(h) $H_z(X)$ with $z = \alpha + jw$ is analytic in all the complex plane except the negative real axis.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

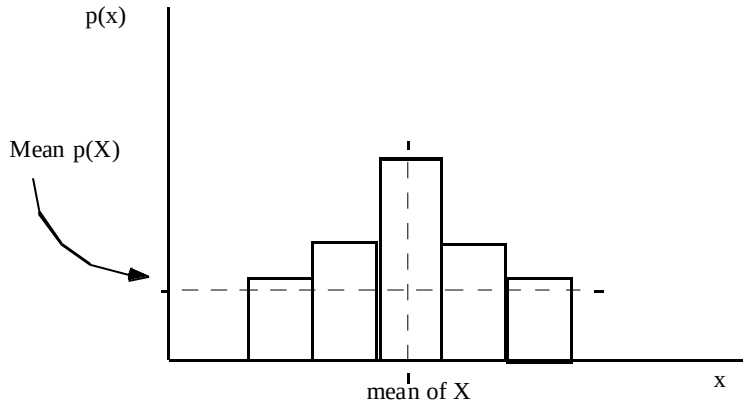
Renyi's quadratic entropy

We will be using heavily Renyi's entropy with α=2, called the quadratic entropy

$$H_2(X) = -\log(\sum_k p_k^2)$$

It has been used in physics, in signal processing and in economics.

We want also to stress that the argument of the log, which is the 2-norm of the PMF $V_2(X)$ has meaning in itself. In fact it is the E[p(X)] or if we make the change in variables ξ_k=p(x_k), then it is the mean of the transformed variable when the PMF is the transformation.



For us Reny's quadratic entropy is appealing because we have found an easy way to estimate it directly from samples.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Extensions to continuous variables.

It is possible to show that Renyi's entropy measure extends to continuous r.v. and reads

$$H_\alpha(X) = \lim_{n\to\infty}(I_\alpha(P_n) - \log n) = \frac{1}{1-\alpha}\log\int p^\alpha(x)dx$$

and for the quadratic entropy

$$H_2(X) = -\log\int p^2(x)dx$$

Notice however that now entropy is no longer positive, in fact it can become arbitrarily large negative.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Estimation of Renyi's quadratic entropy

We will use here ideas from density estimation called kernel density estimation. This is an old method (Rosenblatt) that is now called Parzen density estimation because Parzen proved many important statistical properties of the estimator.

The idea is very simple: Place a kernel over the samples and sum with proper normalization, i.e.

$$\hat{p}_X(x) = \frac{1}{N\sigma}\sum_{i=1}^{N}\kappa\left(\frac{x - x_i}{\sigma}\right)$$

The kernel has the following properties:

1. $\sup_R |k| < \infty$
2. $\int_R |k| < \infty$
3. $\lim_{x\to\infty} |xk(x)| = 0$
4. $k(x) \geq 0, \quad \int_R k(x)dx = 1$

Parzen proved that the estimator is asymptotically unbiased and consistent with a good efficiency.
The free parameter is called the kernel size. Normally a Gaussian is used and σ becomes the standard deviation.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

But notice that here we are not interested in estimating the PDF that is a function, we are interested in a single number the 2-norm of the PDF. For a Gaussian kernel, substituting the estimator yields immediately

$$\hat{H}_2(X) = -\log \int_{-\infty}^{\infty} \left(\frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i) \right)^2 dx$$

$$= -\log \frac{1}{N^2} \int_{-\infty}^{\infty} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} G_\sigma(x - x_j) \cdot G_\sigma(x - x_i) \right) dx$$

$$= -\log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{\infty} G_\sigma(x - x_j) \cdot G_\sigma(x - x_i) dx$$

$$= -\log \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \right)$$

We call this estimator for $V_2(X)$, **the Information Potential.**

Let us look at the derivation: We never had to compute the integral explicitly since the integral of the product of Gaussians is the value of a Gaussian at the difference of arguments (with a larger kernel size).

Let us also look at the expression: The algorithm is $O(N^2)$, and there is a free parameter σ that the user has to select from the data.
Cross validation or the Silverman's rule suffice for most cases

$$\sigma_{opt} = \sigma_X \left(4N^{-1}(2d+1)^{-1} \right)^{\frac{1}{d+4}}$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Extended Estimator for Renyi's Entropy

We can come up with still another estimator for Reny's entropy of any order

$$H_\alpha(X) \stackrel{\Delta}{=} \frac{1}{1-\alpha}\log\int_{-\infty}^{\infty} p_X^\alpha(x)dx = \frac{1}{1-\alpha}\log E_X\left[p_X^{\alpha-1}(X)\right]$$

by approximating the expected value by the empirical mean

$$H_\alpha(X) \approx \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N} p_X^{\alpha-1}(x_j)$$

to yield

$$\hat{H}_\alpha(X) = \frac{1}{1-\alpha}\log\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1}$$

$$= \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_\sigma(x_j - x_i)\right)^{\alpha-1}$$

We will now present a set of properties of these estimators.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Properties of the Estimators

Property 2.1: The estimator of Eq. (2.18) for Renyi's quadratic entropy using Gaussian kernels only differs from the IP of Eq. (2.14) by a factor of $\sqrt{2}$ in the kernel size.

Property 2.2: For any Parzen kernel that obeys the relation

$$\kappa^{new}(x_j - x_i) = \int_{-\infty}^{\infty} \kappa^{old}(x - x_i) \cdot \kappa^{old}(x - x_j) dx \qquad (2.19)$$

the estimator of Renyi's quadratic entropy of Eq. (2.18) matched the estimator of Eq. (2.14) using the IP.

Property 2.3. The kernel size must be a parameter that satisfies the scaling property $\kappa_{c\sigma}(x) = \kappa_{\sigma}(x/c)/c$ for any positive factor c

Property 2.4. The entropy estimator in Eq. (2.18) is invariant to the mean of the underlying density of the samples as is the actual entropy

Property 2.5. The limit of Renyi's entropy as $\alpha \to 1$ is Shannon's entropy. The limit of the entropy estimator in Eq. (2.18) as $\alpha \to 1$ is Shannon's entropy estimated using Parzen windowing with the expectation approximated by the sample mean.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Properties of the Estimators

Property 2.6. In order to maintain consistency with the scaling property of the actual entropy, if the entropy estimate of samples $\{x_1,\ldots,x_N\}$ of a random variable X is estimated using a kernel size of σ, the entropy estimate of the samples $\{cx_1,\ldots,cx_N\}$ of a random variable cX must be estimated using a kernel size of $|c|\sigma$.

Property 2.7. When estimating the joint entropy of an n-dimensional random vector X from its samples $\{x_1,\ldots,x_N\}$, use a multi-dimensional kernel that is the product of single-dimensional kernels. This way, the estimate of the joint entropy and estimate of the marginal entropies are consistent.

Theorem 2.1. The entropy estimator in Eq. (2.18) is consistent if the Parzen windowing and the sample mean are consistent for the actual PDF of the iid samples.

Theorem 2.2. If the maximum value of the kernel $\kappa_\sigma(\xi)$ is achieved when $\xi = 0$, then the minimum value of the entropy estimator in Eq. (2.18) is achieved when all samples are equal to each other, i.e., $x_1 = \ldots = x_N = c$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Properties of the Estimators

Theorem 2.3. If the kernel function $\kappa_\sigma(.)$ is continuous, differentiable, symmetric and unimodal, then the global minimum described in Theorem 2.2 of the entropy estimator in Eq. (2.18) is smooth, i.e., it has a zero gradient and a positive semi-definite Hessian matrix.

Property 2.8. If the kernel function satisfies the conditions in Theorem 2.3, then in the limit, as the kernel size tends to infinity, the quadratic entropy estimator approaches to the logarithm of a scaled and biased version of the sample variance.

Property 2.9. In the case of joint entropy estimation, if the multi-dimensional kernel function satisfies $\kappa_\Sigma(\xi) = \kappa_\Sigma(R^{-1}\xi)$ for all orthonormal matrices, R, then the entropy estimator in Eq. (2.18) is invariant under rotations as is the actual entropy of a random vector X. Notice that the condition on the joint kernel function requires hyper-spherical symmetry.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Properties of the Estimators

Theorem 2.4. $\lim_{N\to\infty} \hat{H}_\alpha(X) = H_\alpha(\hat{X}) \geq H_\alpha(X)$, *where* $\hat{X}$ *is a random variable with the PDF* $f_X(.) * \kappa_\sigma(.)$. *The equality (in the inequality portion) occurs if and only if (iff) the kernel size is zero. This result is also valid on the average for the finite-sample case.*

Bias of the Information Potential

$$Bias[\hat{V}(X)] = E[\hat{V}(X)] - \int f^2(x)dx = (\sigma^2/2)E[f''(X)]$$

Variance of the Information Potential

$$Var(\hat{V}(X)) = E(\hat{V}^2(X)) - (E(\hat{V}(X)))^2 = \frac{aN(N-1)(N-2) + bN(N-1)}{\sigma N^4} \approx \frac{a}{N\sigma}, \quad as \quad N \to \infty$$

$$a = E\left[K_\sigma(x_i - x_j)K_\sigma(x_j - x_l)\right] - E\left[K_\sigma(x_i - x_j)\right]E\left[K_\sigma(x_j - x_l)\right]$$

$$b = Var[K_\sigma(x_i - x_j)]$$

The AMISE (asymptotic mean integrated square error)

$$AMISE(\hat{V}(X)) = E[\int (\hat{V}(X) - V(X))^2 dx] = \frac{\sigma^4}{2}\int (f''(x))^2 dx + \frac{aN(N-1)(N-2) + bN(N-1)}{\sigma N^4}$$

This is expected from the kernel density estimation theory. The IP belongs to this class of estimators.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Physical Interpretation of the Information Potential Estimator

There is a useful physical interpretation of the IP. Think of the gravitational or the electrostatic field.

Assume the samples are particles (information particles) of unit mass that interact with each other with the rules specified by the Parzen kernel used. This analogy is exact!

The information potential field is given by the kernel density estimation

$$\hat{V}_2(x_j) \stackrel{\Delta}{=} \frac{1}{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)$$

the Information Potential is the total potential

$$\hat{V}_2(X) = 1/N \sum_{j=1}^{N} \hat{V}_2(x_j)$$

Now if there is potential there are forces in the space of the samples that can be easily computed as

$$\frac{\partial}{\partial x_j}\hat{V}_2(x_j) = \frac{1}{N}\sum_{i=1}^{N} G'_{\sigma\sqrt{2}}(x_j - x_i) = \frac{1}{2N\sigma^2}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)(x_i - x_j)$$

$$F_2(x_j; x_i) \stackrel{\Delta}{=} \frac{1}{N} G'_{\sigma\sqrt{2}}(x_j - x_i)$$

$$F_2(x_j) \stackrel{\Delta}{=} \frac{\partial}{\partial x_j} V_2(x_j) = \sum_{i=1}^{N} F_2(x_j; x_i)$$

JOSE C. PRINCIPE

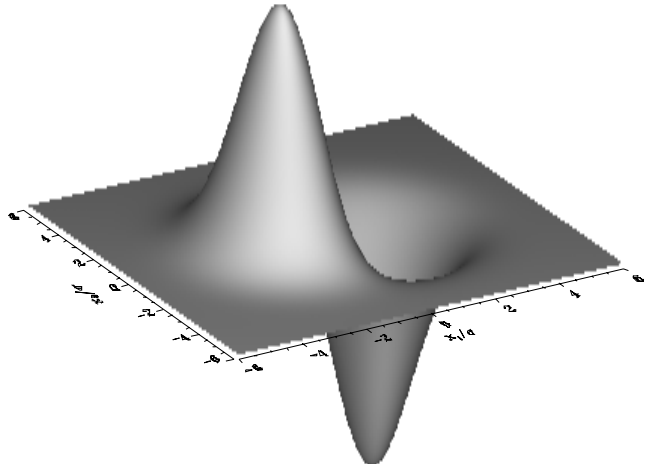
U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Interaction between one information particle in the center of 2-D space



Another way to look at the interactions among samples is to create two matrices in the space

$$\begin{cases} D = \{\mathbf{d}_{ij}\}, & d_{ij} = \mathbf{x}_i - \mathbf{x}_j \\ \varsigma = \{\hat{V}_{ij}\} & \hat{V}_{ij} = G_{\sigma\sqrt{2}}(\mathbf{d}_{ij}) \end{cases} \quad fields \begin{cases} \hat{V}(i) = \dfrac{1}{N}\sum_{j=1}^{N} \hat{V}_{ij} \\ \hat{F}(i) = \dfrac{-1}{N\sigma^2}\sum_{j=1}^{N} \hat{V}_{ij} d_{ij} \end{cases}$$

$$\hat{V}(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{V}_{ij} = \frac{1}{N}\sum_{i=1}^{N} \hat{V}(i)$$

D is a matrix of distances, and ζ is a matrix of scalars. As we can see the samples create a metric space given by the kernel.

JOSE C. PRINCIPE

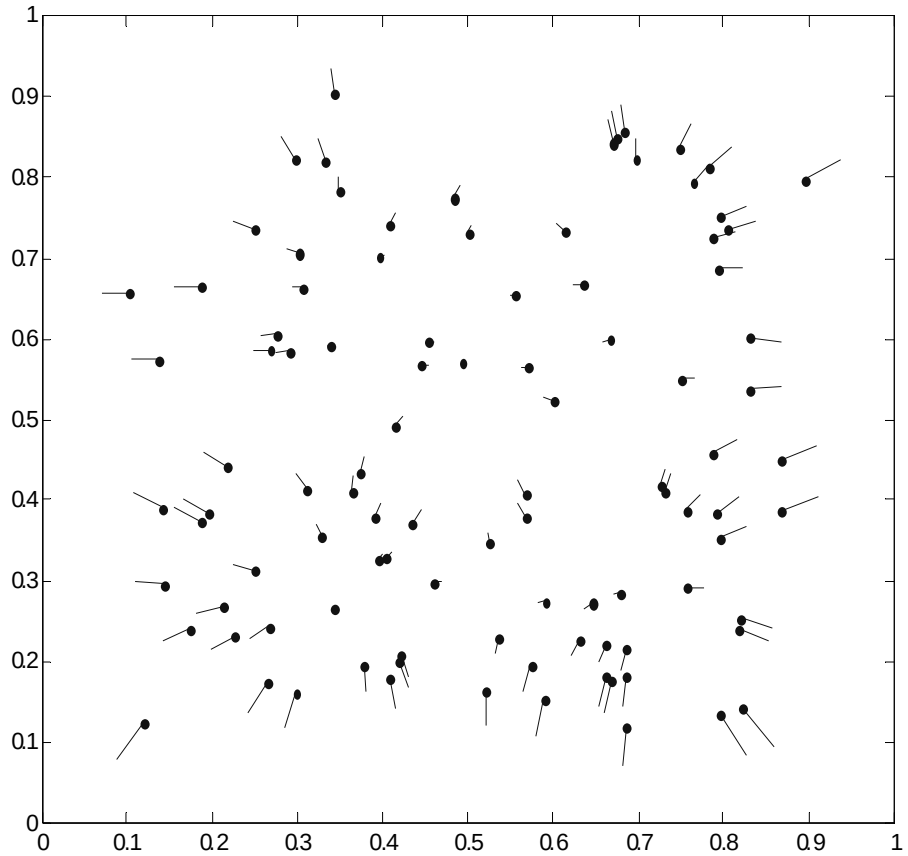
U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Example of forces pulling apart samples when the entropy is maximized. The lines show the direction of the force and their size is the intensity of the pull.



JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Extension to any α and kernel

The framework of information potential and forces extends to any Parzen kernel where the shape of the kernel creates the interaction field. Changing α also changes the field.

The α information potential field is

$$\hat{V}_\alpha(x_j) \stackrel{\Delta}{=} \frac{1}{N^{\alpha-1}} \left(\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right)^{\alpha-1}$$

$$\hat{V}_\alpha(X) = \frac{1}{N} \sum_{j=1}^{N} \hat{V}_\alpha(x_j)$$

and it can be expressed as a function of the IP as

$$\hat{V}_\alpha(x_j) = \frac{1}{N^{\alpha-2}} \left(\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right)^{\alpha-2} \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) = \hat{p}^{\alpha-2}(x_j) \hat{V}_2(x_j)$$

The α information force is likewise given by

$$\hat{F}_\alpha(x_j) \stackrel{\Delta}{=} \frac{\partial}{\partial x_j} \hat{V}_\alpha(x_j) = \frac{\alpha-1}{N^{\alpha-1}} \left(\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i) \right)^{\alpha-2} \left(\sum_{i=1}^{N} \kappa'_\sigma(x_j - x_i) \right)$$

$$= (\alpha-1)\, \hat{p}_X^{\alpha-2}(x_j) \hat{F}_2(x_j)$$

so conceptually all fields can be composed from the IP. $\alpha<2$ emphasizes regions of lower density of samples (opposite for $\alpha>2$).

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Divergence Measures

Renyi proposed a divergence measure from his entropy definition which is different from Kullback Liebler divergence discussed in ch 1.

Renyi's divergence

$$D_\alpha(f \| g) \stackrel{\Delta}{=} \frac{1}{\alpha-1} \log \int_{-\infty}^{\infty} f(x) \left(\frac{f(x)}{g(x)} \right)^{\alpha-1} dx$$

with properties

i. $D_\alpha(f \| g) \geq 0, \ \forall f, g, \alpha > 0$

ii. $D_\alpha(f \| g) = 0$ iff $f(x) = g(x) \ \forall x \in \Re$

iii. $\lim_{\alpha \to 1} D_\alpha(f \| g) = D_{KL}(f \| g)$

$$D_\alpha(f \| g) = \frac{1}{\alpha-1} \log E_p \left[\left(\frac{f(x)}{g(x)} \right)^{\alpha-1} \right] \approx \frac{1}{\alpha-1} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\hat{f}(x_j^f)}{\hat{g}(x_j^g)} \right)^{\alpha-1}$$

$$\frac{1}{\alpha-1} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\sum_{i=1}^{N} \kappa^f(x_j^f - x_i^f)}{\sum_{i=1}^{m} \kappa^g(x_j^g - x_i^g)} \right)^{\alpha-1} = \hat{D}_\alpha(f \| g)$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Reny's Mutual Information

Renyi's mutual information is also the divergence between the joint and the product of the marginals

$$I(X) \stackrel{\Delta}{=} \frac{1}{1-\alpha} \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{p_X^{\alpha}(x^1,...,x^n)}{\prod_{o=1}^{n} p_{X^o}^{1-\alpha}(x^o)} dx^1 ... dx^n$$

and we can come up with a kernel estimate for it if we approximate the E[.] by the empirical mean

$$\hat{I}_{\alpha}(X) \stackrel{\Delta}{=} \frac{1}{1-\alpha} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\left(\frac{1}{N} \sum_{i=1}^{N} \kappa_{\Sigma}(x_j - x_i) \right)}{\prod_{o=1}^{n} \left(\frac{1}{N} \sum_{i=1}^{N} \kappa_{\sigma_o}(x_j^o - x_i^o) \right)} \right)^{1-\alpha}$$

$$= \frac{1}{1-\alpha} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\left(\frac{1}{N} \sum_{i=1}^{N} \prod_{o=1}^{n} \kappa_{\sigma_o}(x_j^o - x_i^o) \right)}{\prod_{o=1}^{n} \left(\frac{1}{N} \sum_{i=1}^{N} \kappa_{\sigma_o}(x_j^o - x_i^o) \right)} \right)^{1-\alpha}$$

that also approximates KL estimator for α->1. However Renyi's divergence and MI are not as general as KL because Shannon measure of information is the only one for which the increase in information is equal to the negative of the decrease in uncertainty.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Quadratic Divergence and Mutual Information

Measuring dissimilarity in probability space is a complex issue and there are many other divergence measures that can be used. If we transform the simplex to an hypersphere preserving the FIsher information, the coordinates of each PMF become $\sqrt{p_k}$. The geodesic distance between PMF f and g becomes $\cos D_G = \sum \sqrt{f_k}\sqrt{g_k}$. This is related to the Bhattacharyya distance (which is Renyi's divergence with α=1/2)

$$D_B(f,g) = -\ln\left(\int \sqrt{f(x)g(x)}dx\right)$$

Now if we measure the distance between f and g in a linear projection space (the cordal distance) we get $D_H = \sum (\sqrt{f_k} - \sqrt{g_k})^{0.5}$ This resembles the Hellinger distance which is also related to Reny's divergence

$$D_H(f,g) = \left[\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx\right]^{1/2} = \left[2\left(1 - \int \sqrt{f(x)g(x)}dx\right)\right]^{1/2}$$

Since we have an estimator for quadratic Renyi's entropy (the IP) we will substitute α=1/2 for 2 and define two quadratic divergences called Euclidean distance and Cauchy Schwarz divergence between f and g.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Euclidean distance between PDFs

We will drop the square root and define the distance as

$$D_{ED}(f,g) = \int (f(x) - g(x))^2 dx = \int f(x)^2 dx - 2\int f(x)g(x)dx + \int g(x)^2 dx$$

Notice that each of the three terms can be estimated with the IP. The middle term involves samples from the two PDFs and so will be called the *Cross Information Potential* (CIP) that measures the potential created by one PDF in the locations specified by the samples of the other PDF.

We will define the quadratic mutual information QMI_{ED} between two random variables as

$$I_{ED}(X_1, X_2) = D_{ED}(f_{X_1X_2}(x_1, x_2), f_{X_1}(x_1) f_{X_2}(x_2))$$

Notice that the I_{ED} is zero if the two r.v. are independent.

The extension to multidimensional random variables is also possible

$$I_{ED}(X_1, ..., X_k) = D_{ED}\left(f_X(x_1, ...x_k), \prod_{i=1}^{k} f_{X_i}(x_i)\right)$$

Although a distance, we sometimes refer to D_{ED} as a divergence.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Cauchy Schwarz divergence

The other divergence is related to the Bhattacharyya distance but can be formally derived from the Cauchy Schwarz inequality

$$\sqrt{\int f^2(x)dx \int g^2(x)dx} \geq \int f(x)g(x)dx$$

where the equality holds iff f(x)=g(x). The divergence is defined as

$$D_{CS}(f,g) = -\log \frac{\int f(x)g(x)dx}{\sqrt{\int f^2(x)dx \int g^2(x)dx}}$$

D_{CS} is always greater than zero, it is zero for f(x)=g(x) and it is symmetric (but does not obey the triangular inequality). We will also work with the square of this quantity in practice. We can also write

$$D_{CS}(f,g) = \log(\int f(x)^2 dx) + \log(\int g(x)^2 dx) - 2\log(\int f(x)g(x)dx)$$

We can also define the quadratic mutual information QMI_{CS} as

$$I_{CS}(X_1, X_2) = D_{CS}(\quad f_{X_1X_2}(x_1, x_2), \quad f_{X_1}(x_1) f_{X_2}(x_2) \quad)$$

and as before for independent variables it is zero. It can also be extended to multiple variables easily

$$I_{CS}(X_1,..., X_k) = D_{CS}\left(f_X(x_1,...x_k), \prod_{i=1}^{k} f_{X_i}(x_i)\right)$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Relation between D_{CS} and Renyi's relative entropy

There is a very interesting interpretation of the D_{CS} in terms of Renyi's entropy. Lutwak defines Renyi's relative entropy as

$$D_{R_\alpha}(f,g) = \log \frac{\left(\int_R g^{\alpha-1}(x) f(x)\right)^{\frac{1}{1-\alpha}} \left(\int_R g^{\alpha}(x)\right)^{1/\alpha}}{\left(\int_R f^{\alpha}(x)\right)^{\frac{1}{\alpha(1-\alpha)}}}$$

for α=2 this gives exactly D_{CS}, so

$$D_{CS}(X,Y) = \log(\int f(x)g(x)dx) - \log(1/2\int f(x)^2 dx) - \log(1/2\int g(x)^2 dx) =$$
$$= H_2(X;Y) - 1/2 H_2(X) - 1/2 H_2(Y)$$

where the first term can be shown to be quadratics Reny's cross entropy.

There is a striking similarity between D_{CS} and Shannon's mutual information, but notice that here we are dealing with Renyi's quadratic entropy.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Geometric Interpretation of Quadratic Mutual Information

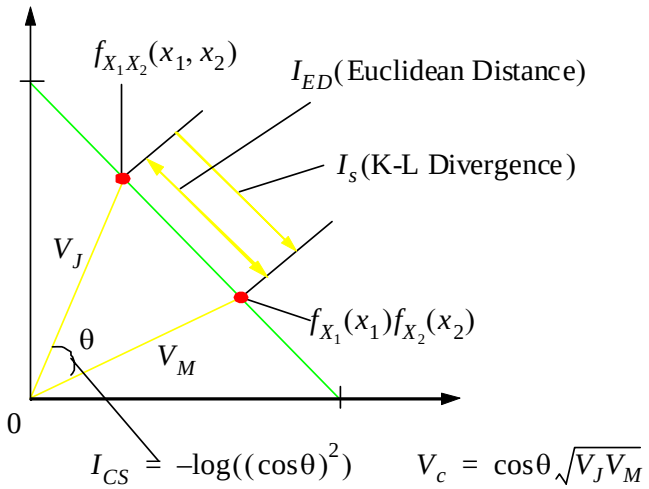
Let us define

$$\begin{cases} V_J = \iint f_{X_1X_2}(x_1, x_2)^2 dx_1 dx_2 \\ V_M = \iint (f_{X_1}(x_1) f_{X_2}(x_2))^2 dx_1 dx_2 \\ V_c = \iint f_{X_1X_2}(x_1, x_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \end{cases}$$

therefore QMI_{ED} and QMI_{ED} can be rewritten as

$$\begin{cases} I_{ED} = V_J - 2V_c + V_M \\ I_{CS} = \log V_J - 2\log V_c + \log V_M \end{cases}$$

The Figure illustrates these distances in the simplex



JOSE C. PRINCIPE

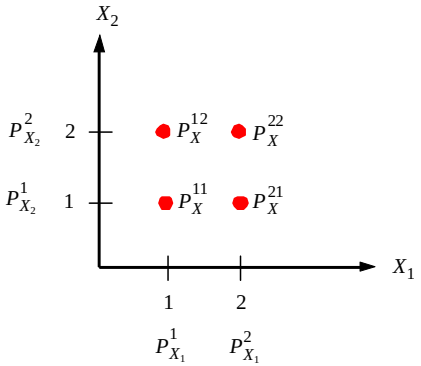
U. OF FLORIDA

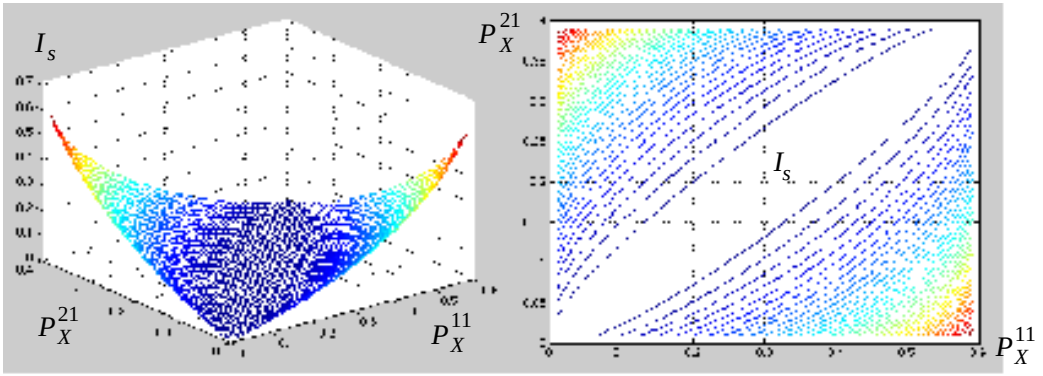
EEL 6935

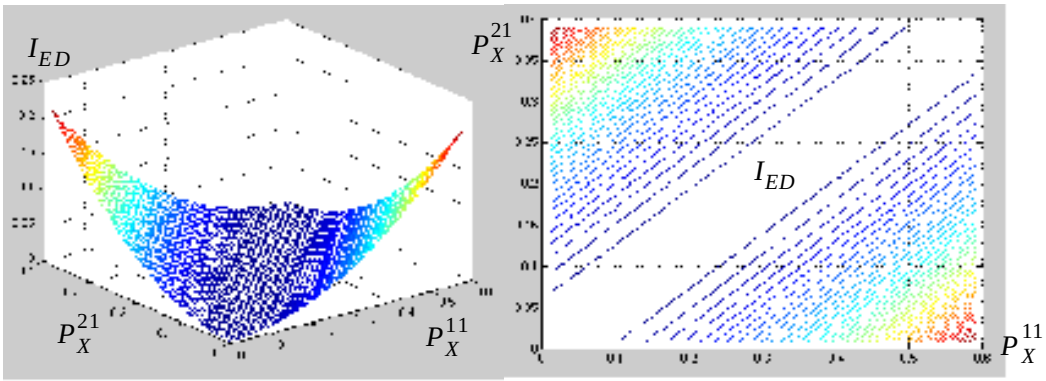
352-392-2662

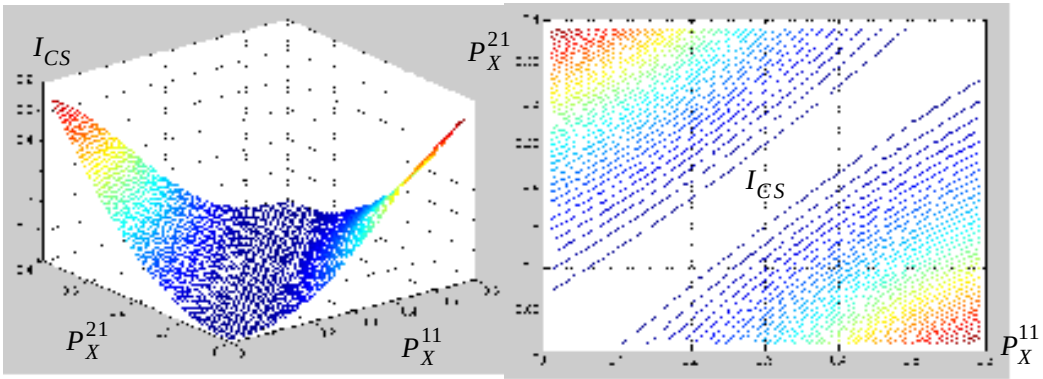
principe@cnel.ufl.edu

Example









Fix P_{X1}= (0.6,0.4)

Vary
P(1,1) -> [0, 0.6]
P(2,1) -> [0, 0.4]

$$I_{ED}(X_1, X_2) = \sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j) - P_{X_1}(i)P_{X_2}(j))^2$$

$$I_{CS}(X_1, X_2) = \log \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j))^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(P_{X1}(i)P_{X_2}(j))^2\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m}(P_X(i,j)P_{X1}(i)P_{X_2}(j))^2}$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Information Potential and Forces in the Joint Spaces

The important lesson to be learned from the IP framework and the distances measures defined, is that each PDF creates its own potential that interacts with the others. This is an additive process, so the forces are also additive, and it becomes relatively simple.

Euclidean and Cauchy Schwarz divergences

The three potentials needed are

$$\hat{V}_f = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_f(i)-x_f(j))^2$$

$$\hat{V}_g = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_g(i)-x_g(j))^2$$

$$\hat{V}_c = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} G_{\sqrt{2}\sigma}(x_f(i)-x_g(j))^2$$

and the divergences and forces are written as

$$\hat{D}_{ED}(f,g) = \hat{V}_{ED} = \hat{V}_f + \hat{V}_g - 2\hat{V}_c \qquad \frac{\partial \hat{V}_{ED}}{\partial x_i} = \frac{\partial \hat{V}_f}{\partial x_i} + \frac{\partial \hat{V}_g}{\partial x_i} - 2\frac{\partial \hat{V}_c}{\partial x_i}$$

$$\hat{D}_{CS}(f,g) = \hat{V}_{CS} = \log\frac{\hat{V}_f \hat{V}_g}{\hat{V}_c^{\,2}} \qquad \frac{\partial \hat{V}_{CS}}{\partial x_i} = \frac{1}{\hat{V}_f}\frac{\partial \hat{V}_f}{\partial x_i} + \frac{1}{\hat{V}_g}\frac{\partial \hat{V}_g}{\partial x_i} - \frac{2}{\hat{V}_c}\frac{\partial \hat{V}_c}{\partial x_i}$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Quadratic Mutual Information

The QMIs are a bit more detailed because now we have two deal with the joints, the marginals and their product. But everything is still additive. The three PDFs are

$$\begin{cases} \hat{f}_{X_1X_2}(x_1,x_2) = \dfrac{1}{N}\sum_{i=1}^{N} G_\sigma(\boldsymbol{x}-\boldsymbol{x}(i)) \\ \hat{f}_{X_1}(x_1) = \dfrac{1}{N}\sum_{i=1}^{N} G_\sigma(x_1-x_1(i)) \\ \hat{f}_{X_2}(x_2) = \dfrac{1}{N}\sum_{i=1}^{N} G_\sigma(x_2-x_2(i)) \end{cases}$$

they create the following fields

$$\hat{V}_J = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(\boldsymbol{x}(i)-\boldsymbol{x}(j)) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i)-x_1(j))G_{\sqrt{2}\sigma}(x_2(i)-x_2(j))$$

$$\hat{V}_M = \hat{V}_1\hat{V}_2 \qquad with \quad \hat{V}_k = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_k(i)-x_k(j)), \qquad k=1,2$$

$$\hat{V}_C = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i)-x_1(j))\right)\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(i)-x_2(j))\right) \qquad (2.104)$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

let us exemplify for V_C

$$\hat{V}_C = \int\int \hat{f}(x_1, x_2)\hat{f}(x_1)\hat{f}(x_2)dx_1dx_2$$

$$= \int\int \left[\frac{1}{N}\sum_{k=1}^{N} G_\sigma(x_1 - x_1(k))G_\sigma(x_2 - x_2(k))\right]\left[\frac{1}{N}\sum_{i=1}^{N} G_\sigma(x_1 - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_\sigma(x_2 - x_2(j))\right]dx_1dx_2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{N}\sum_{k=1}^{N}\int G_\sigma(x_1 - x_1(i))G_\sigma(x_1 - x_1(k))dx_1 \int G_\sigma(x_2 - x_2(k))G_\sigma(x_2 - x_2(j))dx_2$$

$$= \frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sqrt{2}\sigma}(x_1(k) - x_1(i))\right]\left[\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(k) - x_2(j))\right] \qquad (2.102)$$

Notice that Vc has complexity O(N^3). Finally we have

$$\hat{I}_{ED}(X_1, X_2) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))G_{\sqrt{2}\sigma}(x_2(i) - x_2(j)) + \hat{V}_1\hat{V}_2 -$$

$$- \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))\right)\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(i) - x_2(j))\right)$$

$$\hat{I}_{CS}(X_1, X_2) = \log \frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))G_{\sqrt{2}\sigma}(x_2(i) - x_2(j))\right)\left(\hat{V}_1\hat{V}_2\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_1(i) - x_1(j))\right)\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(x_2(i) - x_2(j))\right)\right)^2}$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

The interactions are always based on the marginals, but there are three levels: the level of the individual marginal samples (joint space), the level of the marginal sample and marginal fields (the CIP) and the product of the two marginal fields. This field can be generalized to any number of variables

$$\hat{I}_{ED}(X_1,...,X_K) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{k=1}^{K}\hat{V}_k(ij) - \frac{2}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\hat{V}_k(i) + \prod_{k=1}^{K}\hat{V}_k$$

$$\hat{I}_{CS}(X_1,...,X_K) = \log\frac{\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\prod_{k=1}^{K}\hat{V}_k(ij)\right)\prod_{k=1}^{K}\hat{V}_k}{\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\hat{V}_k(i)\right)^2}$$

The information forces for each field are also easily derived

$$\hat{F}_{ED}(i) = \frac{\partial\hat{I}_{CS}}{\partial x_k(i)} = \frac{\partial\hat{V}_j}{\partial x_k(i)} - \frac{2\partial\hat{V}_C}{\partial x_k(i)} + \frac{\partial\hat{V}_k}{\partial x_k(i)}$$

$$\hat{F}_{CS}(i) = \frac{\partial\hat{I}_{CS}}{\partial x_k(i)} = \frac{1}{V_j}\frac{\partial\hat{V}_j}{\partial x_k(i)} - \frac{2}{V_C}\frac{\partial\hat{V}_C}{\partial x_k(i)} + \frac{1}{V_k}\frac{\partial\hat{V}_k}{\partial x_k(i)}$$

These expressions are going to be very useful when adapting systems with divergences or quadratic mutual information.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Fast Information and Cross Information Potential Calculations

One of the problems with this IP estimator methodology is the computational complexity which is O(N^2) or O(N^3). There are two ways to get approximations to the IP and CIP with arbitrary accuracy: The Fast Gauss Transform and the Incomplete Cholesky decomposition. They both transform the complexity to O(NM) where M <<N.

Fast Gauss Transform

The FGT takes advantage of the fast decay of the Gaussian function and can efficiently compute weighted sums of scalar Gaussians.

$$S(y_i) = \sum_{j=1}^{N} w_i e^{-(y_j - y_i)^2 / 4\sigma^2} \qquad i = 1,...M$$

The savings come from the shifting property of the Gaussian function which reads

$$e^{-\left(\frac{y_j - y_i}{\sigma}\right)^2} = e^{-\left(\frac{y_j - y_C - (y_i - y_C)}{\sigma}\right)^2} = e^{-\left(\frac{y_j - y_c}{\sigma}\right)^2} \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{y_i - y_c}{\sigma}\right)^n H_n\left(\frac{y_j - y_c}{\sigma}\right) \qquad h_n(y) = (-1)^n \frac{d^n \exp(-x^2)}{dx^n}$$

and the efficient approximation of the Hermite polynomials with a small order p.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

The shifting property is useful because there is no need to evaluate every Gaussian at every point. Instead a p term sum is computed around a small number y_c of cluster centers with O(Np) computation. These sums are then shifted to the y_i desired locations and computed in another O(Mp) operations. Normally the centers c are found using the furthest point algorithm which is efficient.

The information potential calculation (M=N) can be immediately given as

$$V(y) \approx \frac{1}{2\sigma N^2\sqrt{\pi}} \sum_{j=1}^{N} \sum_{b=1}^{B} \sum_{n=0}^{p-1} \frac{1}{n!} h_n\left(\frac{y_j - y_{C_b}}{2\sigma}\right) C_n(b) \qquad C_n(b) = \sum_{y_i \in B} \left(\frac{y_j - y_{C_b}}{2\sigma}\right)^n$$

which requires O(NpB) calculations. p is normally 4 or 5 (independent of N), and B, the number of clusters is also relatively small (~10), but with a weak dependence on N.

The problem with this algorithm is if the data is multidimensional because the complexity p changes to p^D where D is the dimension. In order to cope with high dimensions, a vector Taylor series has been proposed to approximate the high dimensional Gaussian as

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

$$\exp\left(-\frac{\left\|\mathbf{y}_j-\mathbf{y}_i\right\|^2}{4\sigma^2}\right)=\exp\left(-\frac{\left\|\mathbf{y}_j-c\right\|^2}{4\sigma^2}\right)\exp\left(-\frac{\left\|\mathbf{y}_i-c\right\|^2}{4\sigma^2}\right)\exp\left(2\frac{(\mathbf{y}_j-c)\cdot(\mathbf{y}_i-c)}{4\sigma^2}\right)$$

and the cross term is approximated by a Taylor expansion as

$$\exp\left(2\frac{(\mathbf{y}_j-c)\cdot(\mathbf{y}_i-c)}{4\sigma^2}\right)=\sum_{\alpha\geq 0}\frac{2^{|\alpha|}}{\alpha!}\left(\frac{\mathbf{y}_j-c}{2\sigma}\right)^{\alpha}\left(\frac{\mathbf{y}_i-c}{2\sigma}\right)^{\alpha}+\varepsilon(\alpha)$$

$$\alpha!=\alpha_1!\,\alpha_2!\cdots\alpha_d!$$

$$|\alpha|=\alpha_1+\alpha_2+\cdots+\alpha_d$$

Now the IP for multidimensions becomes

$$V_T(\mathbf{y})\approx\frac{1}{N^2\left(4\pi\sigma^2\right)^{d/2}}\sum_{j=1}^{N}\sum_{B}\sum_{\alpha\geq 0}C_\alpha(B)\exp\left(-\frac{\left\|\mathbf{y}_j-c_B\right\|^2}{4\sigma^2}\right)\left(\frac{\mathbf{y}_j-c_B}{2\sigma}\right)^{\alpha}$$

$$C_\alpha(B)=\frac{2^{|\alpha|}}{\alpha!}\left\{\sum_{e_i\in B}\exp\left(-\frac{\left\|\mathbf{y}_i-c_B\right\|^2}{4\sigma^2}\right)\left(\frac{\mathbf{y}_i-c_B}{2\sigma}\right)^{\alpha}\right\}$$

For a D dimensional data set the calculation is $O(NBr_{p,D})$ instead of $O(NBr^D)$ with $r=\binom{p+D}{d}=\frac{(p+D)!}{D!p!}$, but normally p must be larger than before for the same precision.

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

Incomplete Cholesky Decomposition

It turns out that the eigenvalues of the matrix created by the pairwise evaluation of the Gaussian is full rank, but the eigenspectrum decays very fast. Hence we can take advantage of this fact.

For a NxN symmetric matrix $K=G^T G$, where G is a lower triangular matrix of positive diagonal entries. When the eigenspectum falls fast we can approximate K by NxD lower triangular matrix $\tilde{G}$ such that $\left\| K - \tilde{G}^T \tilde{G} \right\| < \varepsilon$. There are ways of computing $\tilde{G}$ effectively. The computation complexity of the procedure becomes O(ND).

The IP can be computed as

$$\hat{V}(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} k(x_i - x_j) = \frac{1}{N^2} \mathbf{1}_N^T K_{XX} \mathbf{1}_N = \frac{1}{N^2} \left\| \mathbf{1}_N^T \tilde{G}_{XX} \right\|_2^2$$

$$\hat{I}_{ED} = \frac{1}{N^2} \mathbf{1}_{\mathbf{D}_x}^{\mathbf{T}} (\tilde{G}^T{}_{XX} \tilde{G}_{YY} \circ \tilde{G}^T{}_{XX} \tilde{G}_{YY}) \mathbf{1}_{\mathbf{D}_y} + \frac{1}{N^4} \left\| \mathbf{1}_N^T \tilde{G}_{XX} \right\|_2^2 \left\| \mathbf{1}_N^T \tilde{G}_{YY} \right\|_2^2 - \frac{2}{N^3} (\mathbf{1}_N^T \tilde{G}_{XX})(\tilde{G}^T{}_{XX} \tilde{G}_{YY})(\tilde{G}_{YY}^T \mathbf{1}_N)$$

$$\hat{I}_{CS} = \log \frac{\mathbf{1}_{\mathbf{D}_x}^{\mathbf{T}} (\tilde{G}^T{}_{XX} \tilde{G}_{YY} \circ \tilde{G}^T{}_{XX} \tilde{G}_{YY}) \mathbf{1}_{\mathbf{D}_y} \left\| \mathbf{1}_N^T \tilde{G}_{XX} \right\|_2^2 \left\| \mathbf{1}_N^T \tilde{G}_{YY} \right\|_2^2}{\left((\mathbf{1}_N^T \tilde{G}_{XX})(\tilde{G}_{XX}^T \tilde{G}_{YY})(\tilde{G}_{YY}^T \mathbf{1}_N) \right)^2}$$

JOSE C. PRINCIPE

U. OF FLORIDA

EEL 6935

352-392-2662

principe@cnel.ufl.edu

However the CIP is not a positive definite matrix, so we need to extend K_{XX} to create a positive definite matrix $K_{ZZ} = \begin{vmatrix} K_{XX} & K_{XY} \\ K_{XY} & K_{YY} \end{vmatrix}$ where K_{XY} denotes the Gram matrix for the CIP. The calculation of the CIP becomes

$$\hat{V}(X;Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i - y_j) = \frac{1}{N^2}\mathbf{e}_1^T K_{zz}\mathbf{e}_2 = \frac{1}{N^2}\left(\mathbf{e}_1^T \tilde{G}_{ZZ}\right)\left(\tilde{G}_{ZZ}\mathbf{e}_2\right)$$

$$\mathbf{e}_1 = \{\underbrace{1,\ldots 1}_{N}, \underbrace{0,\ldots 0}_{N}\}^T$$

$$\mathbf{e}_2 = \{\underbrace{0,\ldots 0}_{N}, \underbrace{1,\ldots 1}_{N}\}^T$$

with complexity $O(ND^2)$. The divergence measures can be computed efficiently as

$$\hat{D}_{ED} = \frac{1}{N^2}(\mathbf{e_1^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_1}) + \frac{1}{N^2}(\mathbf{e_2^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_2}) - \frac{2}{N^2}(\mathbf{e_1^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_2})$$

$$\hat{D}_{CS} = \log\frac{(\mathbf{e_1^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_1})(\mathbf{e_2^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_2})}{\left((\mathbf{e_1^T}\tilde{G}_{ZZ})(\tilde{G}_{ZZ}^T\mathbf{e_2})\right)^2}$$