Information Theoretic Learning

Jose C. Principe

Computational NeuroEngineering Laboratory
Electrical and Computer Engineering Department
University of Florida
www.cnel.ufl.edu
principe@cnel.ufl.edu

Acknowledgments

Dr. John Fisher
Dr. Dongxin Xu
Dr. Ken Hild
Dr. Deniz Erdogmus
Dr. Puskal Pokharel
Dr. Weifeng Liu
Dr. Jianwu Xu
Dr. Kyu-Hwa Jeong
Dr. Sudhir Rao
Dr. Seungju Han

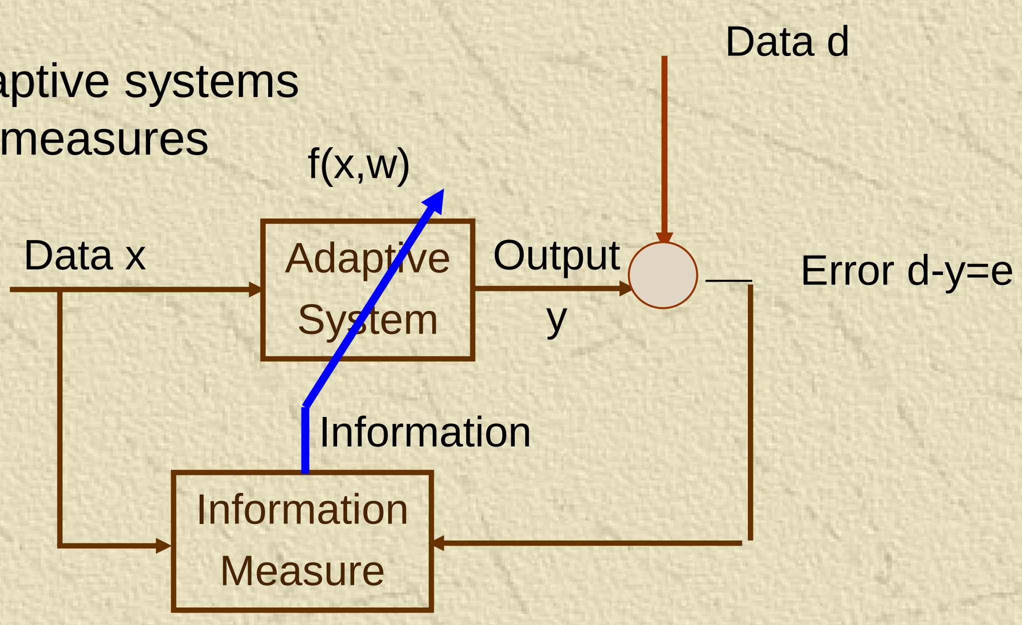
NSF ECS – 0300340, 0601271, 0856441
(Neuroengineering program) and DARPA

Outline

- Information Theoretic Learning
- A RKHS for ITL
- Correntropy as Generalized Correlation
- How do we learn from the environment? The principle of relevant information

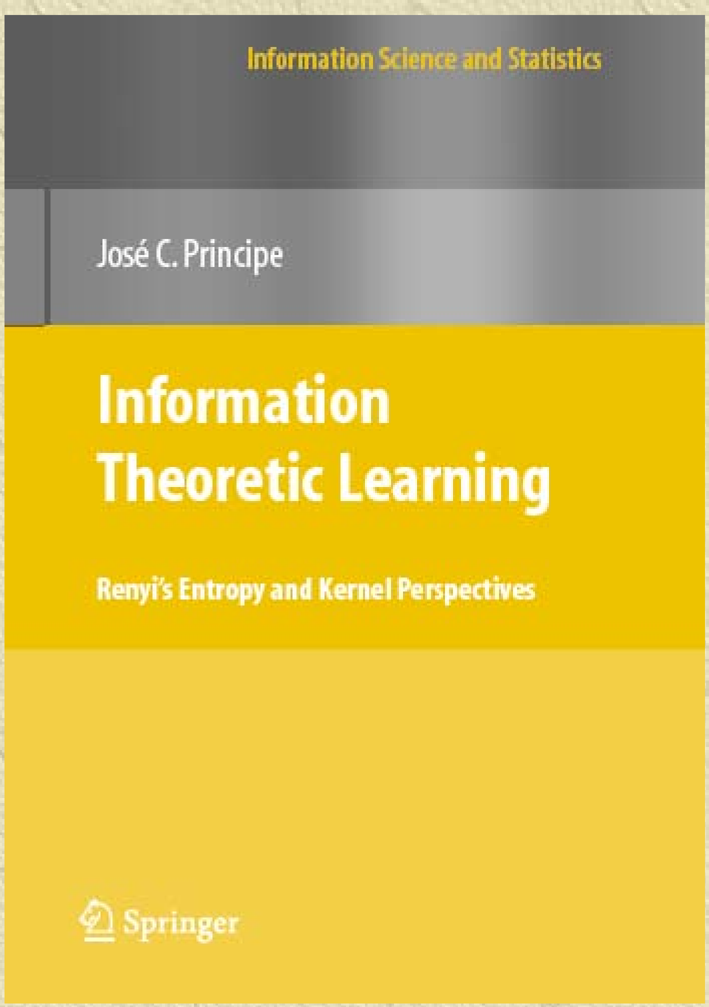
Information Filtering: From Data to Information

- Information Filters: Given data pairs $\{x_i,d_i\}$
  - Optimal Adaptive systems
  - Information measures



- Embed information in the weights of the adaptive system
- More formally, use optimization to perform Bayesian estimation

Information Theoretic Learning (ITL)-2010



Tutorial

IEEE
SP MAGAZINE, Nov 2006

Or ITL resource
www.cnel.ufl.edu

What is Information Theoretic Learning?

ITL is a methodology to adapt linear or nonlinear systems using criteria based on the information descriptors of entropy and divergence.

Center piece is a non-parametric estimator for entropy that:

- Does not require an explicit estimation of pdf
- Uses the Parzen window method which is known to be consistent and efficient
- Estimator is smooth
- Readily integrated in conventional gradient descent learning
- Provides a link between information theory and Kernel learning.

ITL is a different way of thinking about data quantification

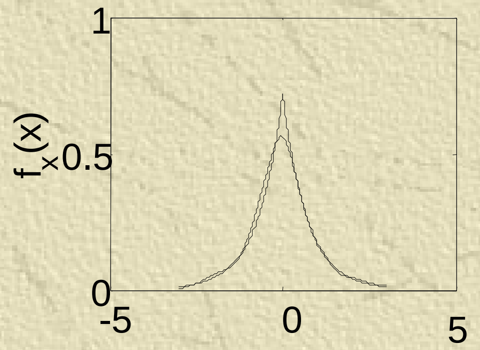
Moment expansions, in particular **Second Order moments** are still today the workhorse of statistics. We automatically translate deep concepts (e.g. similarity, Hebb's postulate of learning) in 2^{nd} order statistical equivalents.

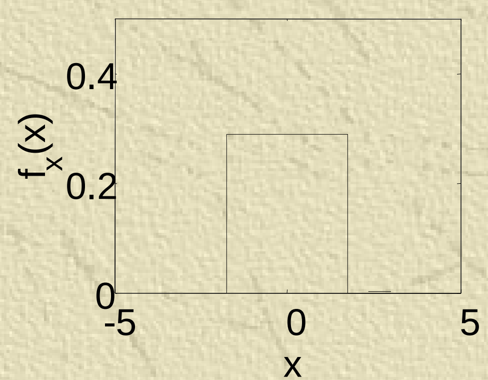
ITL replaces 2^{nd} order moments with a geometric statistical interpretation of data in probability spaces.

- Variance by **Entropy**
- Correlation by **Correntopy**
- Mean square error (MSE) by **Minimum error entropy** (MEE)
- Distances in data space by distances in probability spaces
- Fully exploits the strucutre of RKHS.

Information Theoretic Learning
Entropy

Not all random variables (r.v.) are equally random!





- Entropy quantifies the degree of uncertainty in a r.v. Claude Shannon defined entropy as

$$H_S(X) = -\sum p_X(x) \log p_X(x) \qquad H_S(X) = -\int f_X(x) \log(f_X(x))dx$$

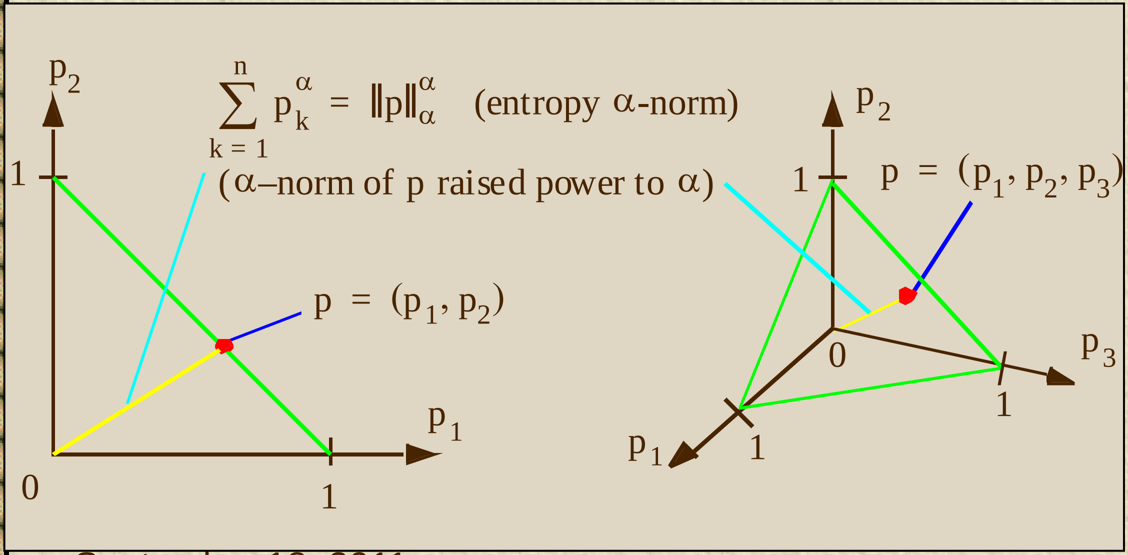
Information Theoretic Learning

Renyi's Entropy

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\sum p_X^\alpha(x) \qquad H_\alpha(X) = \frac{1}{1-\alpha}\log\int f_X^\alpha(x)dx$$

Renyi's entropy equals Shannon's as $\alpha \to 1$

- Norm of the pdf:



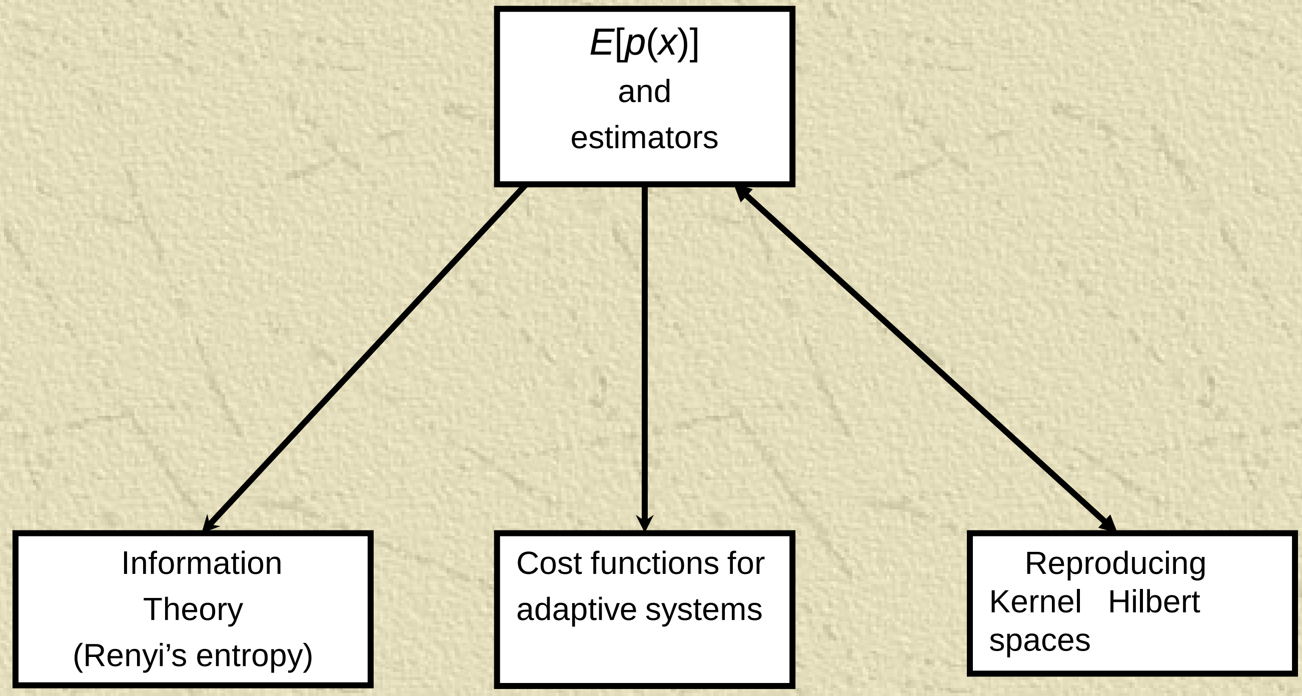
$$V_\alpha = \int f^\alpha(x)dx$$

$$\alpha - norm = \sqrt[\alpha]{V_\alpha}$$

V_α=α-Information Potential

Information Theoretic Learning
Norm of the PDF (Information Potential)

$V_2(x)$, 2- norm of the pdf (Information Potential) is one of the central concept in ITL.



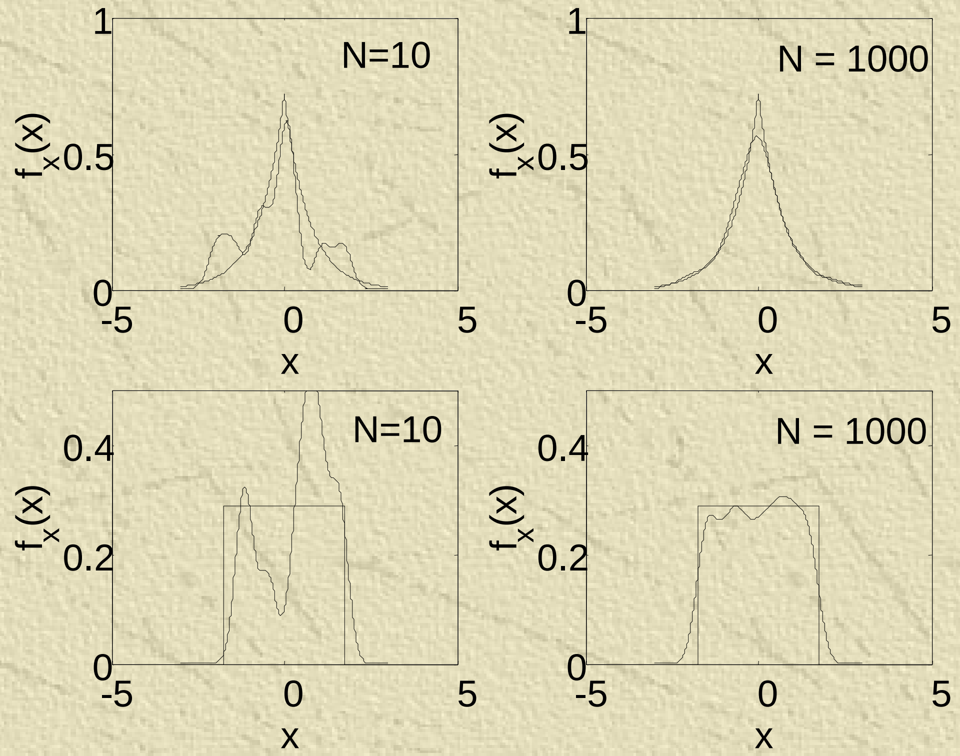
Information Theoretic Learning
Parzen windowing

Given <u>only</u> samples drawn from a distribution:

$$\{x_1,\ldots,x_N\} \sim p(x)$$

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(x - x_i)$$

Kernel function



Convergence:

$$\lim_{N\to\infty} \hat{p}(x) = p(x) * G_{\sigma(N)}(x)$$

$$provided\ that\ N\sigma(N) \to \infty$$

Information Theoretic Learning
Information Potential

Order-2 entropy & Gaussian kernels:

$$H_2(X) = -\log \int p^2(x)dx = -\log \int \left(\frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i) \right)^2 dx$$

$$= -\log \left(\frac{1}{N^2} \sum_j \sum_i \int G_\sigma(x - x_j) G_\sigma(x - x_i) dx \right)$$

$$= -\log \underbrace{\left(\frac{1}{N^2} \sum_j \sum_i G_{\sigma\sqrt{2}}(x_j - x_i) \right)}$$

Pairwise interactions between samples $O(N^2)$

Information potential estimator, $\hat{V}_2(X)$

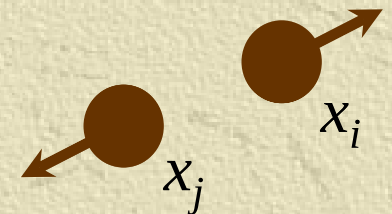
$\hat{p}(x)$ provides a potential field over the space of the samples parameterized by the kernel size σ

Principe, Fisher, Xu, *Unsupervised Adaptive Filtering*, (S. Haykin), Wiley, 2000.

Information Theoretic Learning

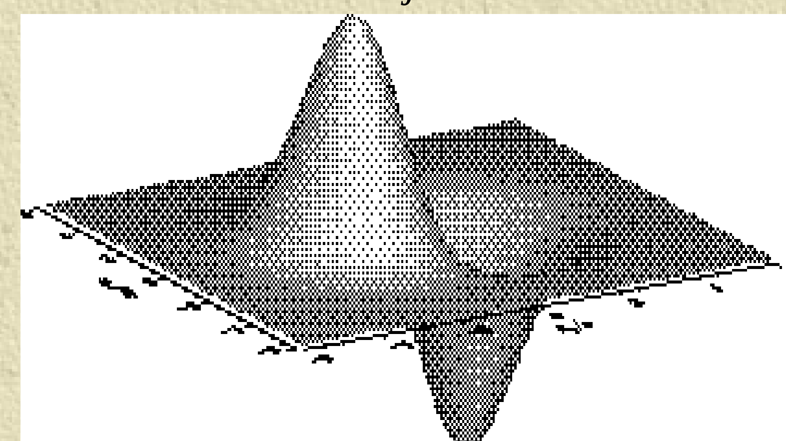
Information Force

- In adaptation, samples become *information particles* that interact through information forces.



Information potential field:

$$\hat{V}_2(x_j) = \frac{1}{N}\sum_i G_{\sigma\sqrt{2}}(x_j - x_i)$$



Information force:

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

Principe, Fisher, Xu, *Unsupervised Adaptive Filtering*, (S. Haykin), Wiley, 2000.

Erdogmus, Principe, Hild, Natural Computing, 2002.

Information Theoretic Learning

Error Entropy Criterion (EEC)

We will use iterative algorithms for optimization of a linear system with steepest descent

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \eta \nabla \mathbf{V_2}(n)$$

Given a batch of N samples the IP is

$$\hat{V}_2(E) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(e_i - e_j)$$

For an FIR the gradient becomes

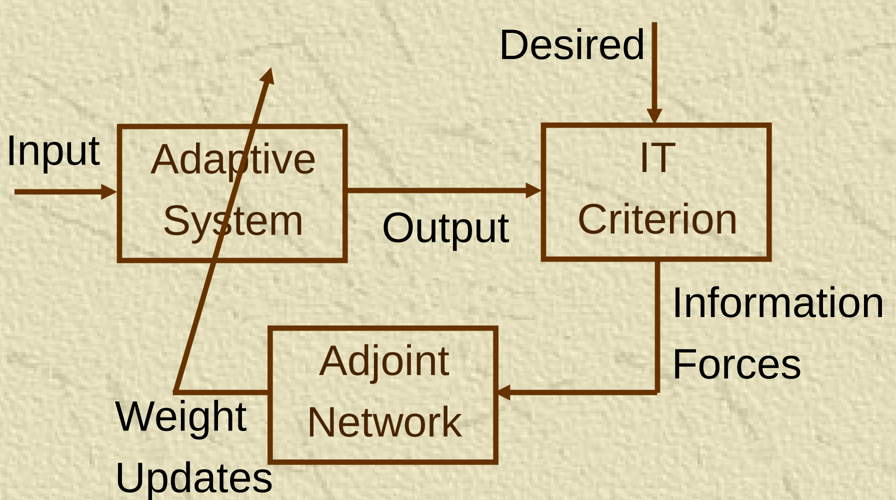
$$\nabla_k \hat{V}_2(n) = \frac{\partial \hat{V}(e(n))}{\partial w_k} = \frac{\partial\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}\sigma}(e_i - e_j)\right)}{\partial(e(n-i) - e(n-j))} \frac{\partial(e(n-i) - e(n-j))}{\partial w_k} =$$

$$\frac{2}{N}\sum_{i=1}^{N}\ \sum_{j=1}^{N} G_\sigma(e(n-i) - e(n-j))(e(n-i) - e(n-j))\left(\frac{\partial y(n-j)}{\partial w_k} - \frac{\partial y(n-i)}{\partial w_k}\right)]$$

$$\nabla_k \hat{V}_2(n) = \frac{2}{N^2}\sum_{i=1}^{N}\ \sum_{j=1}^{N} G_\sigma(e(n-i) - e(n-j))(e(n-i) - e(n-j))(x_k(n-j) - x_k(n-i))$$

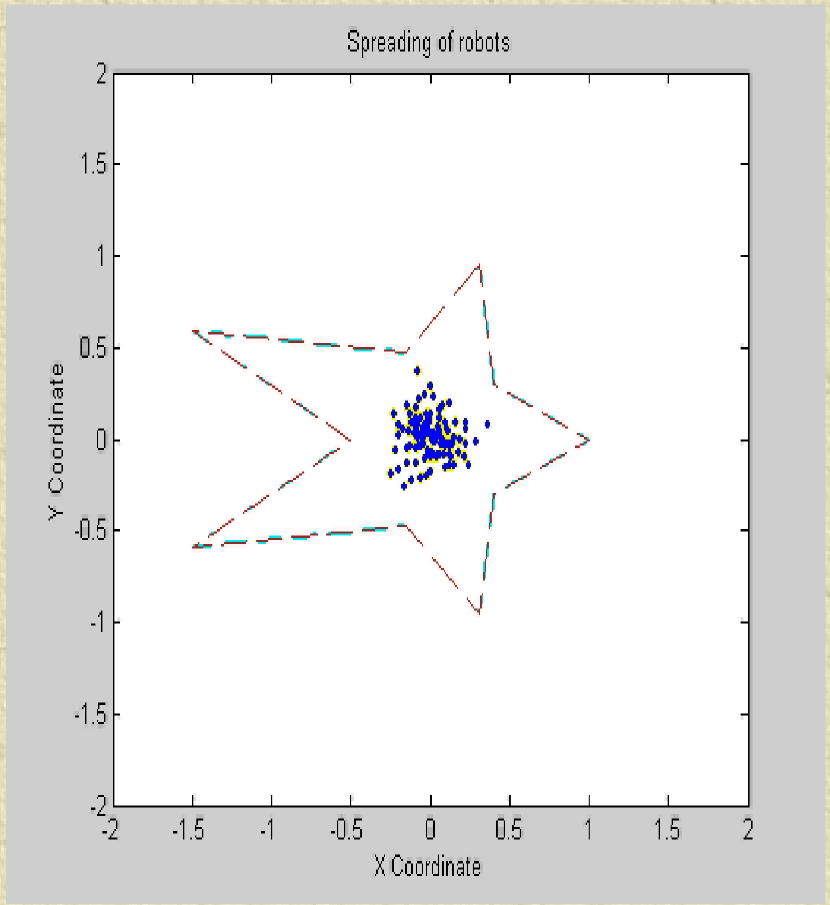
Information Theoretic Learning

Backpropagation of Information Forces



$$\frac{\partial J}{\partial w_{ij}} = \sum_{p=1}^{k}\sum_{n=1}^{N} \frac{\partial J}{\partial e_p(n)} \frac{\partial e_p(n)}{\partial w_{ij}}$$

Information forces become the **injected error** to the dual or adjoint network that determines the weight updates for adaptation.



Information Theoretic Learning

Quadratic divergence measures

Kulback-Liebler Divergence:

$$D_{KL}(X;Y) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Renyi's Divergence:

$$D_{\alpha}(X;Y) = \frac{1}{\alpha - 1} \log \int p(x) \left(\frac{p(x)}{q(x)} \right)^{\alpha - 1} dx$$

Euclidean Distance:

$$D_E(X;Y) = \int \left(p(x) - q(x) \right)^2 dx$$

Cauchy- Schwartz Divergence :

$$D_C(X;Y) = -\log\left(\frac{\int p(x) q(x) dx}{\sqrt{\int p^2(x) dx \int q^2(x) dx}} \right)$$

Mutual Information is a special case (distance between the joint and the product of marginals)

Information Theoretic Learning
How to estimate Euclidean Distance

Euclidean Distance: $D_E(p;q) = \int \left(p(x) - q(x)\right)^2 dx$

$$D_E(p;q) = \int p^2(x)dx - 2\int p(x)q(x)dx + \int q^2(x)dx =$$

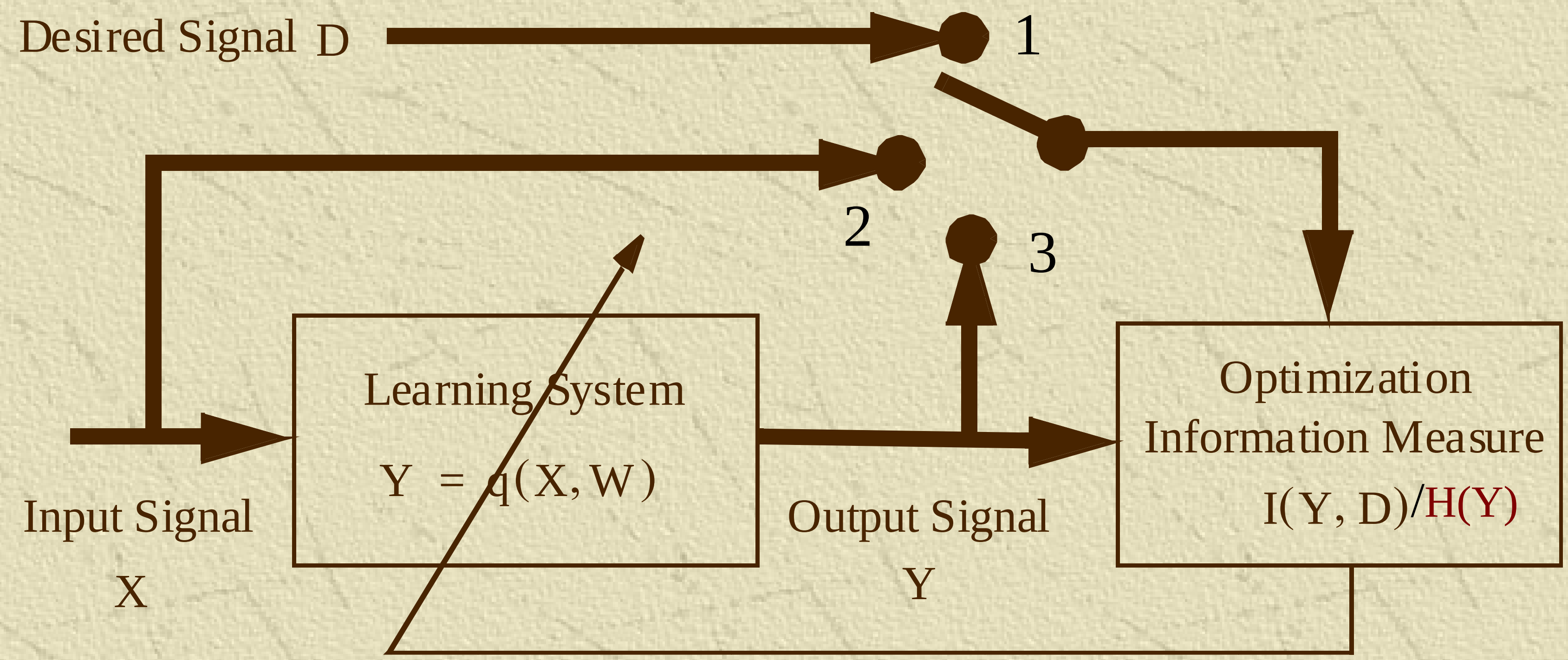
$$= \frac{1}{N_p N_p}\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} G_{\sigma\sqrt{2}}(x_i - x_j) - \frac{2}{N_p N_q}\sum_{i=1}^{N_p}\sum_{a=1}^{N_q} G_{\sigma\sqrt{2}}(x_i - x_a) + \frac{1}{N_q N_q}\sum_{a=1}^{N_q}\sum_{b=1}^{N_q} G_{\sigma\sqrt{2}}(x_a - x_b)$$

$\int p(x)q(x)dx$ is called the **cross information potential (CIP)**

So D_{ED} can be readily computed with the information potential . Likewise for the Cauchy Schwartz divergence, and also the quadratic mutual information

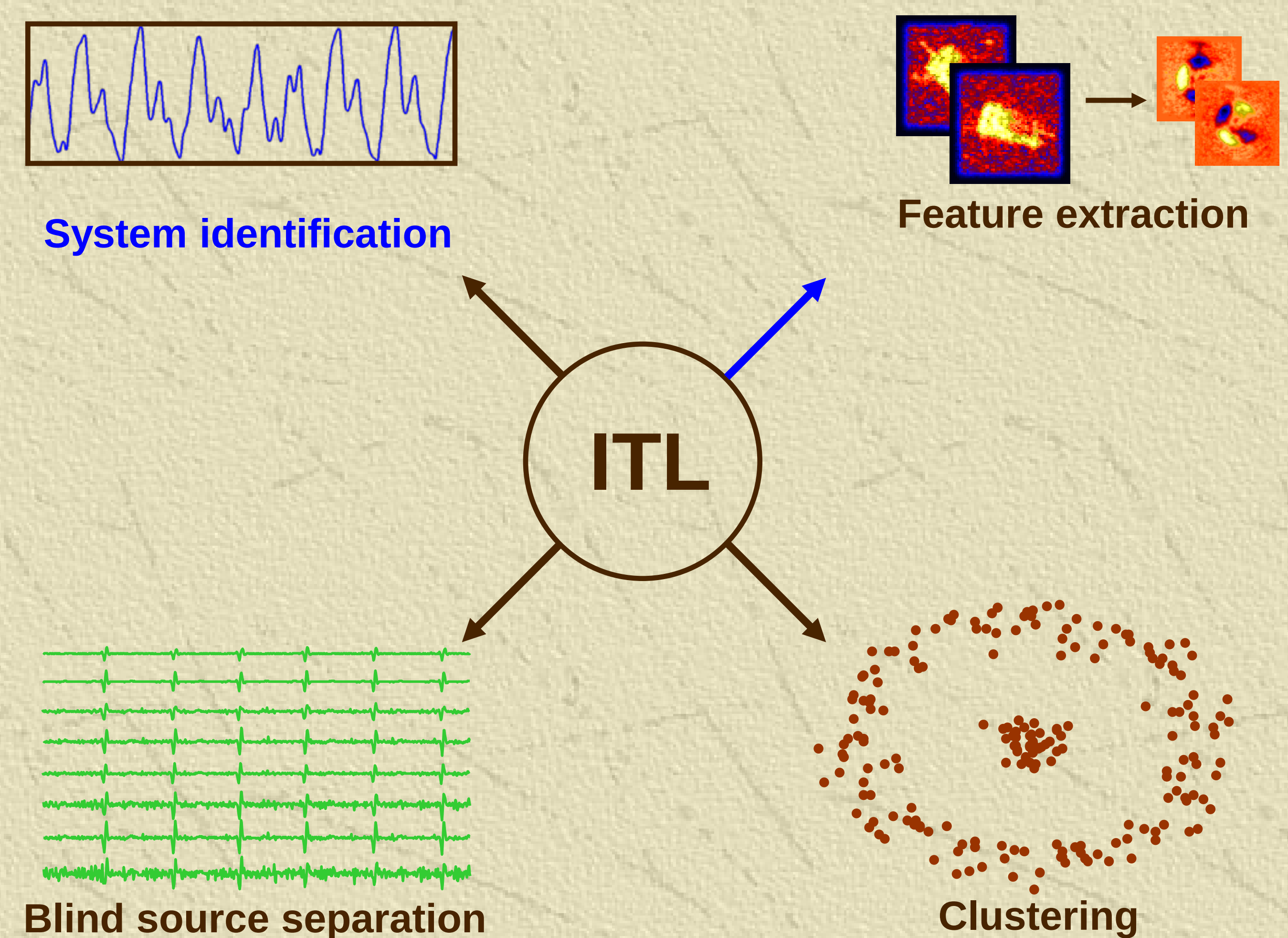
Principe, Fisher, Xu, *Unsupervised Adaptive Filtering*, (S. Haykin), Wiley, 2000.

Information Theoretic Learning
Unifies supervised and unsupervised learning



Switch 1	Switch 2	Switch 3
Filtering/classification	InfoMax	ICA
Feature extraction		Clustering
(also with entropy)		NLPCA

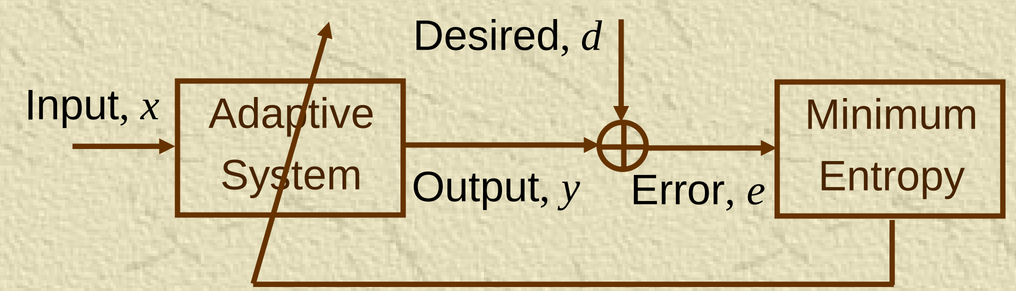
ITL - Applications



www.cnel.ufl.edu → ITL has examples and Matlab code

ITL – Applications
Nonlinear system identification

- Minimize information content of the residual error



- Equivalently provides the best density matching between the output and the desired signals.

$$\min_{\mathbf{w}} \frac{1}{1-\alpha} \log \int p_e^{\alpha}(\varepsilon;\mathbf{w})d\varepsilon \equiv \min_{\mathbf{w}} \iint p_{xy}(\xi,\eta;\mathbf{w}) \left(\frac{p_{xy}(\xi,\eta;\mathbf{w})}{p_{xd}(\xi,\eta)} \right)^{\alpha-1} d\xi d\eta$$

Erdogmus, Principe, IEEE Trans. Signal Processing, 2002. (*IEEE SP 2003 Young Author Award*)

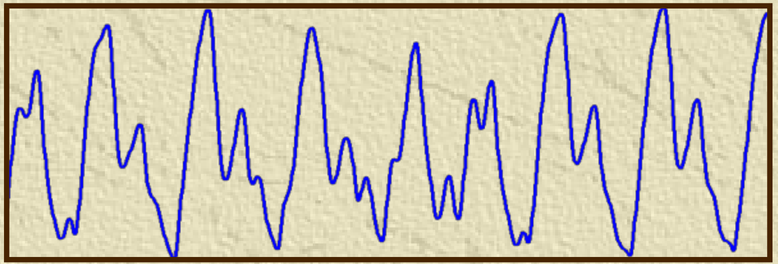
ITL – Applications
Time-series prediction

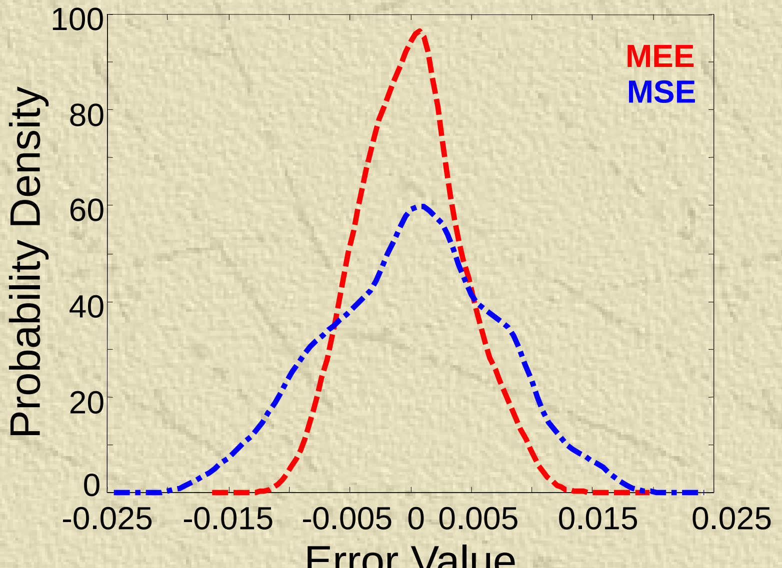
Chaotic Mackey-Glass (MG-30) series

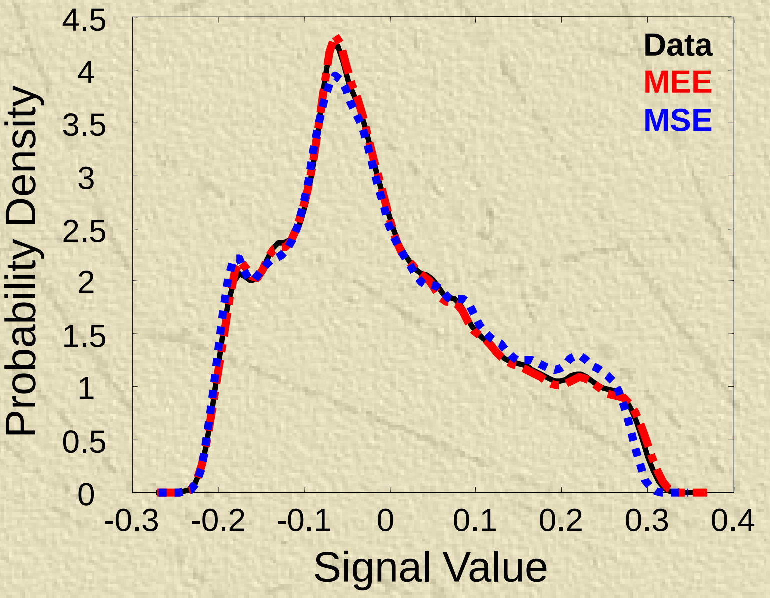
Compare 2 criteria:

Minimum squared-error

Minimum error entropy

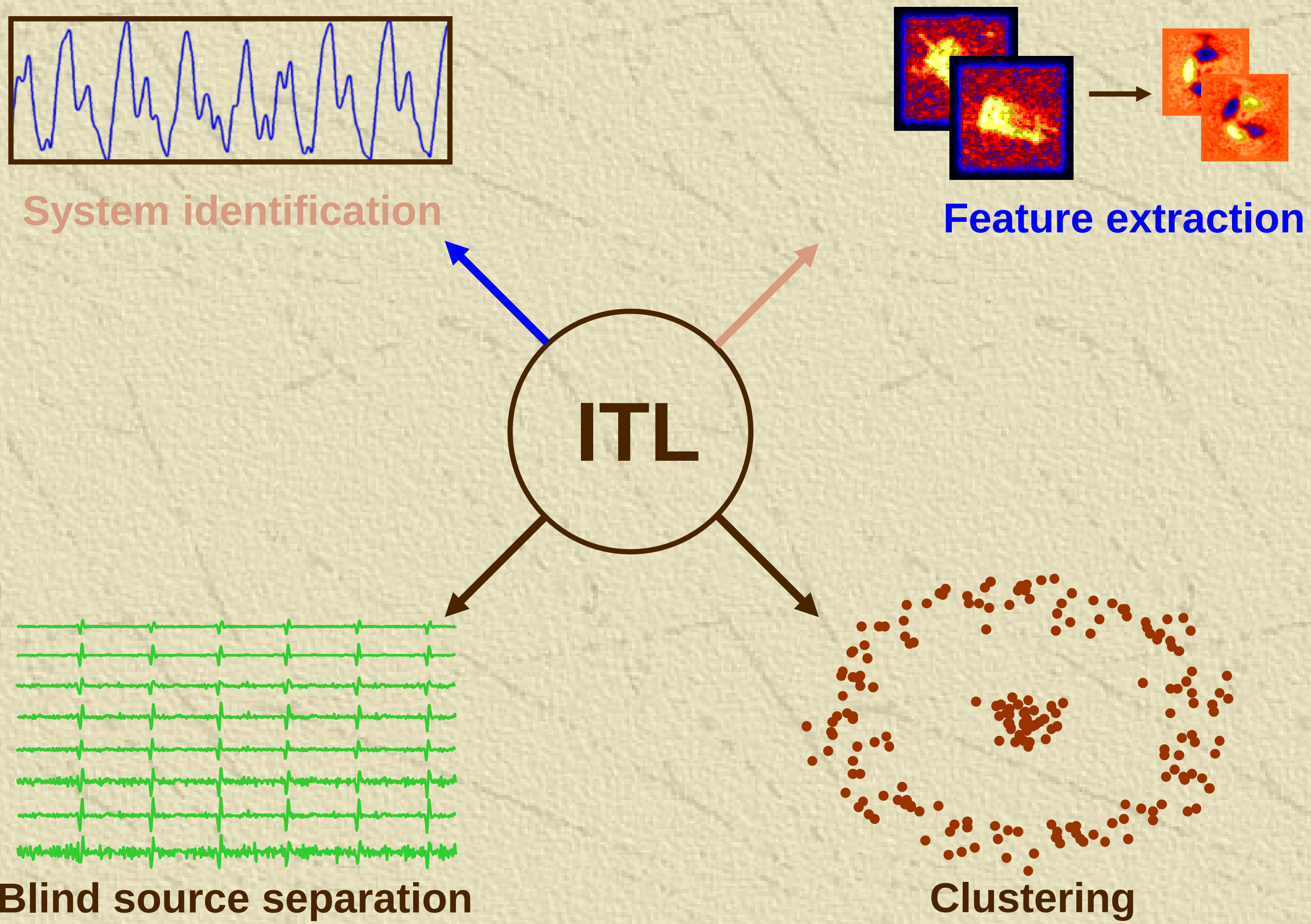






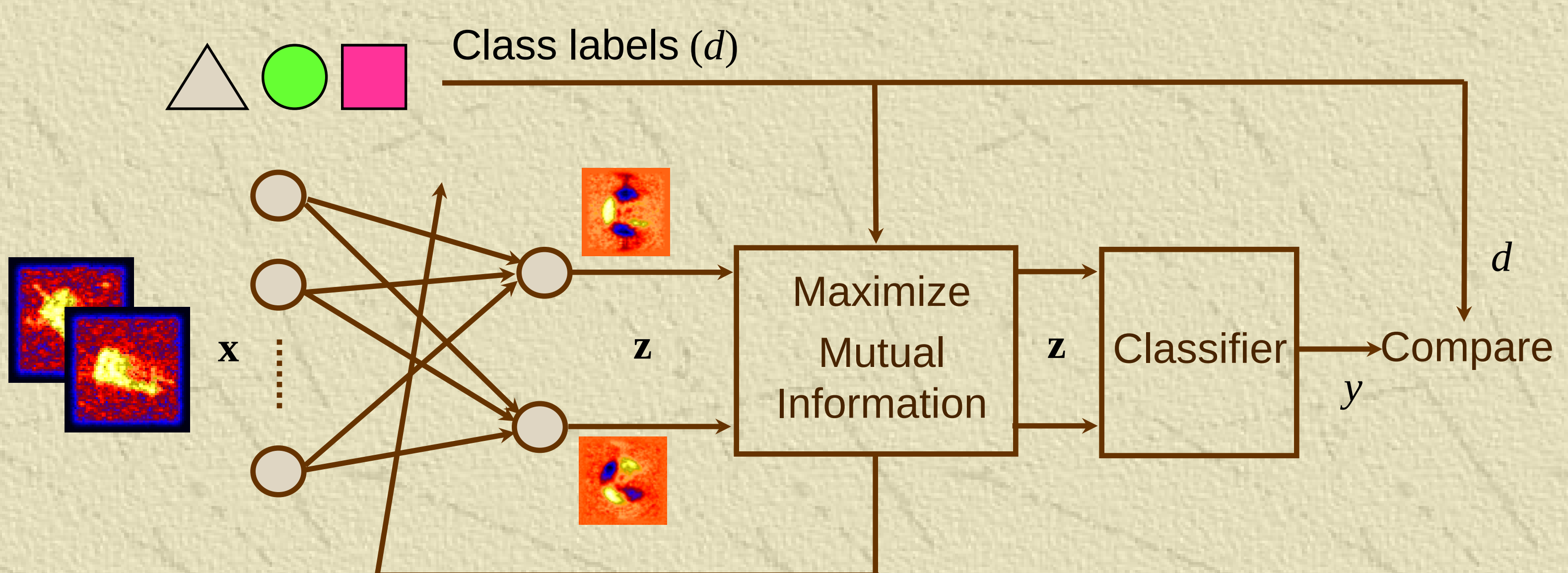
System: TDNN (6+4+1)

ITL - Applications



ITL – Applications
Optimal feature extraction

- Data processing inequality: Mutual information is monotonically non-increasing.
- Classification error inequality: Error probability is bounded from below and above by the mutual information.



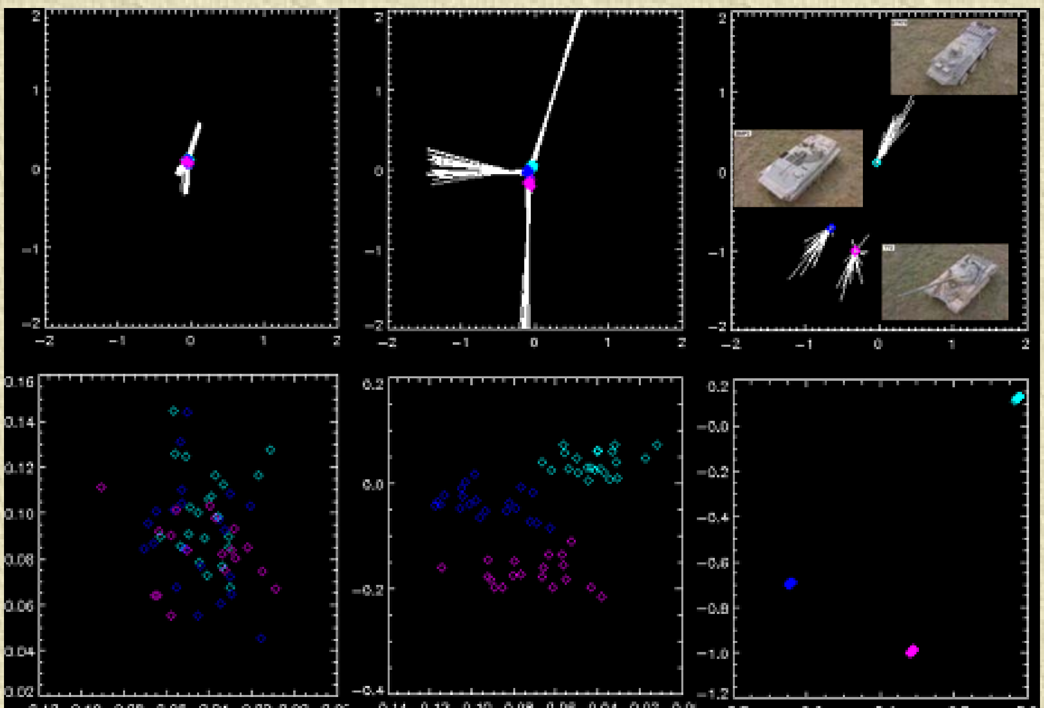
PhD on feature extraction for sonar target recognition (2002)

Principe, Fisher, Xu, *Unsupervised Adaptive Filtering*, (S. Haykin), Wiley, 2000.

ITL – Applications
Extract 2 nonlinear features

64x64 SAR images of 3 vehicles: BMP2, BTR70, T72

Information forces in training



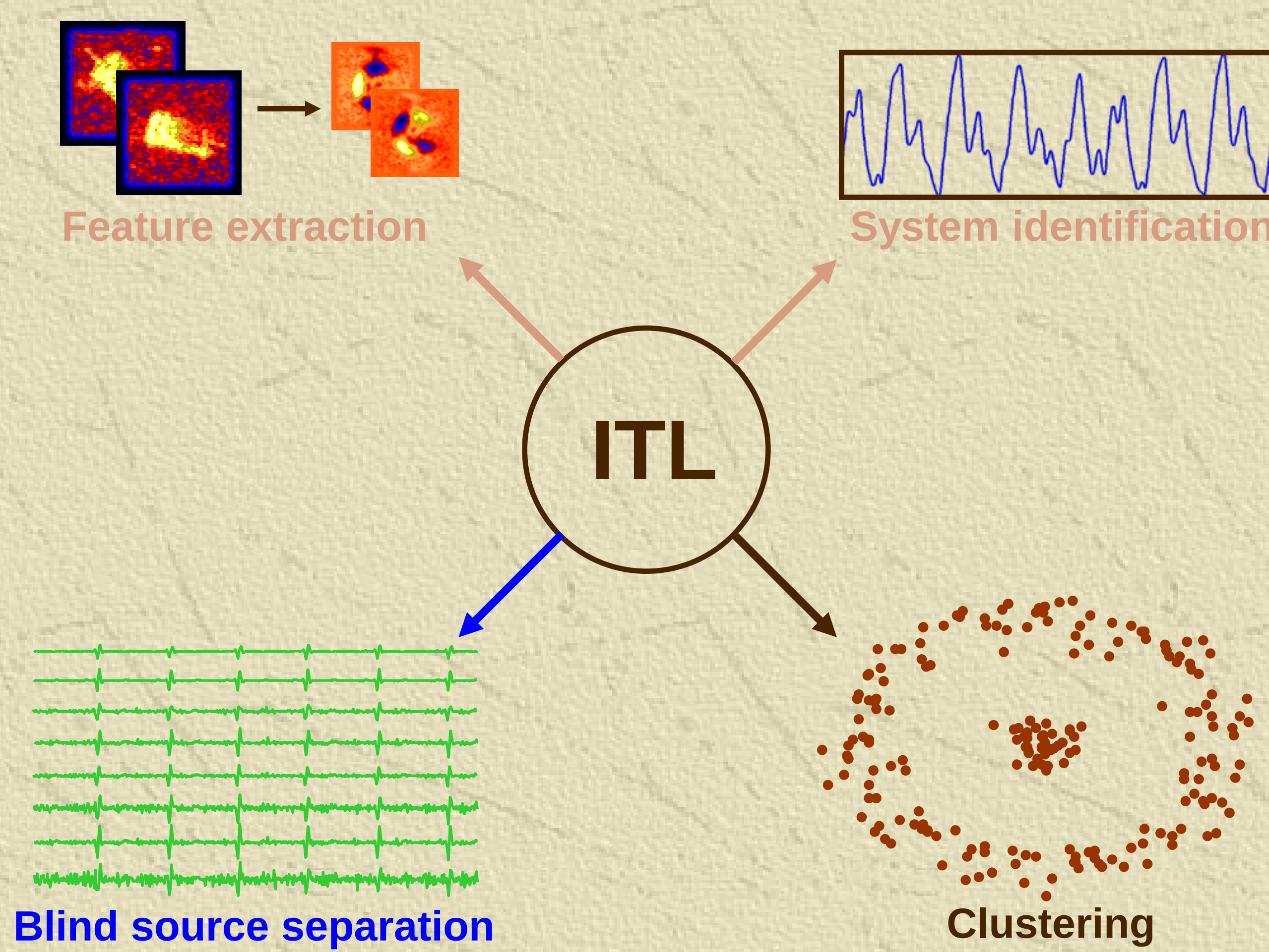
Classification results

	P(Correct)
MI+LR	94.89%
SVM	94.60%
Templates	90.40%

Zhao, Xu and Principe, SPIE Automatic Target Recognition, 1999.

Hild, Erdogmus, Principe, IJCNN Tutorial on ITL, 2003.

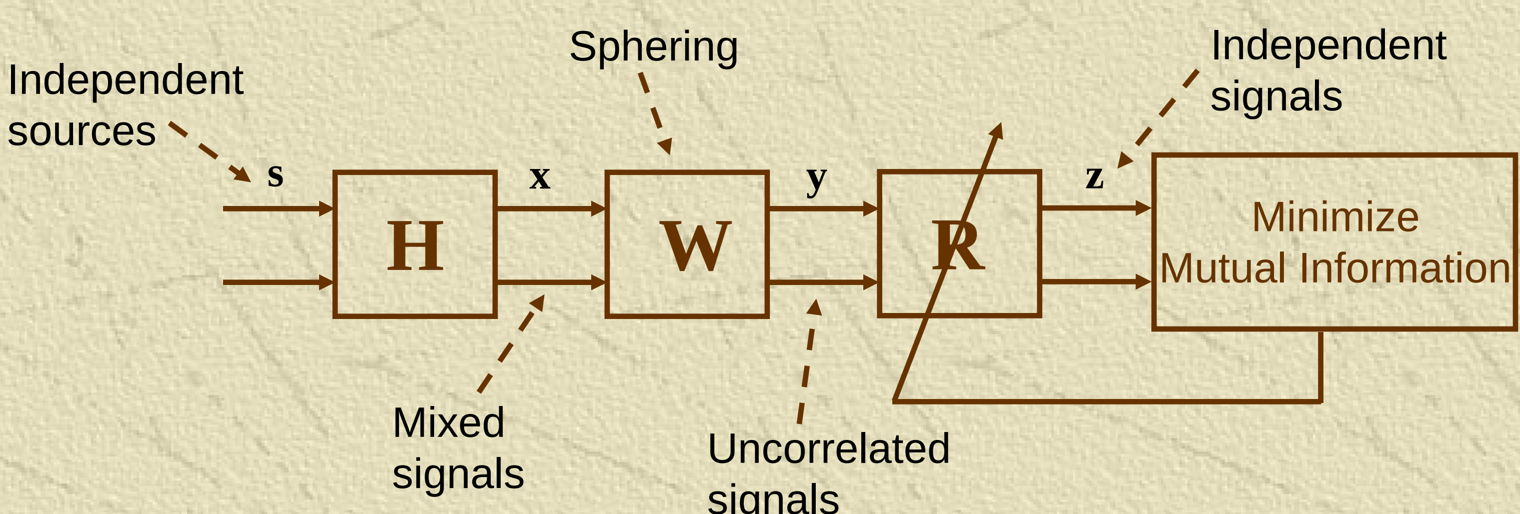
ITL - Applications



ITL – Applications
Independent component analysis

- Observations are generated by an unknown mixture of statistically independent unknown sources.

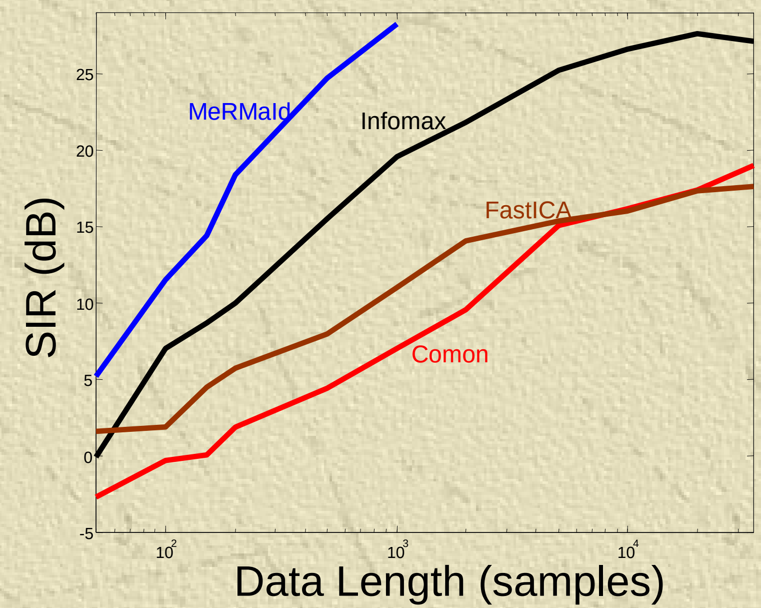
$$\mathbf{x}_k = \mathbf{H}\mathbf{s}_k \qquad I(\mathbf{z}) = \sum_{c=1}^{n} H(z_c) - \cancel{H(\mathbf{z})}$$



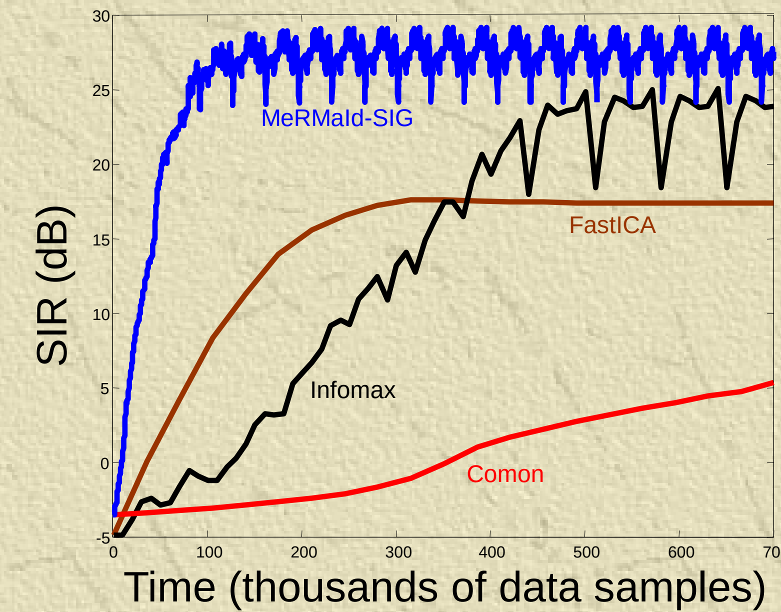
Ken Hild II, PhD on blind source separation (2003)

ITL – Applications
On-line separation of mixed sounds

Off-line separation, 10x10 mixture



On-line separation, 3x3 mixture



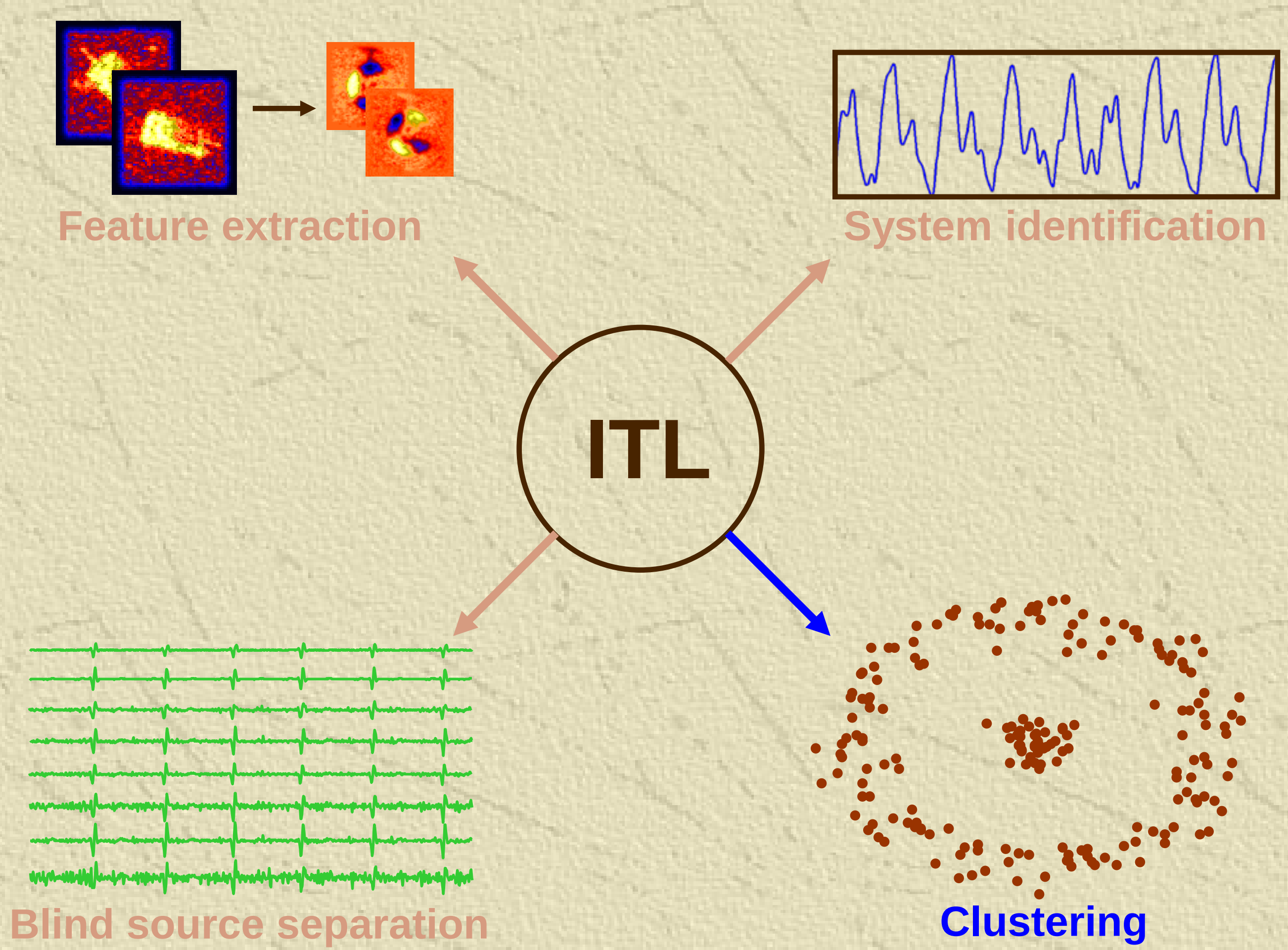
Observed mixtures and separated outputs

X1: X2: X3:

Z1: Z2: Z3:

Hild, Erdogmus, Principe, IEEE Signal Processing Letters, 2001.

ITL - Applications



ITL – Applications
Information theoretic clustering

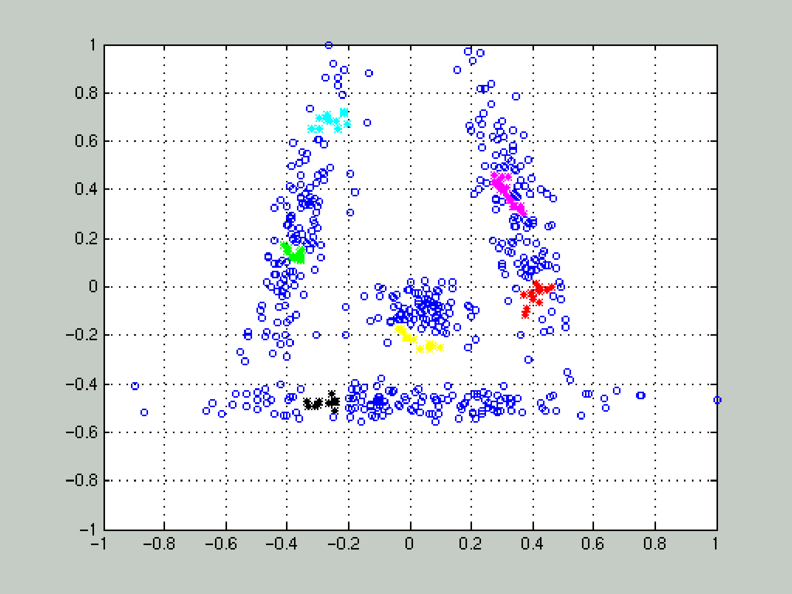
- Select clusters based on entropy and divergence:
  - Minimize within cluster entropy
  - Maximize between cluster divergence

Between cluster divergence

$$\min_{\mathbf{m}} \frac{\int p(x)q(x)dx}{\left(\int p^2(x)dx \int q^2(x)dx\right)^{1/2}}$$

Membership vector

Within cluster entropy



Robert Jenssen PhD on information theoretic clustering

Jenssen, Erdogmus, Hild, Principe, Eltoft, IEEE Trans. Pattern Analysis and Machine Intelligence, 2005. (submitted)

Reproducing Kernel Hilbert Spaces as a Tool for Nonlinear System Analysis

Fundamentals of Kernel Methods

Kernel methods are a very important class of algorithms for nonlinear optimal signal processing and machine learning. Effectively they are <u>shallow</u> (one layer) neural networks (RBFs) for the Gaussian kernel.

- They exploit the linear structure of Reproducing Kernel Hilbert Spaces (RKHS) with very efficient computation.
- ANY (!) SP algorithm expressed in terms of inner products has in principle an equivalent representation in a RKHS, and may correspond to a nonlinear operation in the input space.
- Solutions may be <u>analytic instead of adaptive</u>, when the linear structure is used.

Fundamentals of Kernel Methods

RKHS induced by the Gaussian kernel

The Gaussian kernel is symmetric and positive definite

$$k_\sigma(x, x') = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-x')^2}{2\sigma^2}\right).$$

thus induces a RKHS on a sample set $\{x_1, \ldots x_N\}$ of reals, denoted as $RKHS_K$.

Further, by Mercer's theorem, a kernel mapping Φ can be constructed which transforms data from the input space to $RKHS_K$ where:

$$k_\sigma(x - x_i) = \left\langle \Phi(x), \Phi(x_i) \right\rangle_K$$

where <,> denotes inner product in $RKHS_K$.

A RKHS for ITL

RKHS induced by cross information potential

Let *E* be the set of all square integrable one dimensional probability density functions, i.e., $f_i(x) \in E, \quad \forall i \in I$ where $\int f_i^2(x)dx < \infty$ and I is an index set. Then form a linear manifold (similar to the simplex)

$$\left\{ \sum_i \alpha_i f_i(x) \right\}$$

Close the set and define a proper inner product

$$\left\langle f_i(x), f_j(x) \right\rangle_{L_2} = \int f_i(x) f_j(x) dx$$

$L_2(E)$ is an Hilbert space but it is not reproducing. However, let us define the bivariate function on $L_2(E)$ (**cross information potential** (CIP))

$$V(f_i, f_j) = \int f_i(x) f_j(x) dx$$

One can show that the CIP is a positive definite function and so it defines a $RKHS_V$. Moreover there is a congruence between $L_2(E)$ and H_V.

A RKHS for ITL
ITL cost functions in RKHS$_V$

- Cross Information Potential (is the natural distance in H$_V$)

$$\int f(x)g(x)dx = \left\langle V(f,.), V(g,.)\right\rangle_{H_V}$$

- Information Potential (is the norm (mean) square in H$_V$)

$$\int f(x)f(x)dx = \left\langle V(f,.), V(f,.)\right\rangle_{H_V} = \left\| V(f,.)\right\|^2$$

- Second order statistics in H$_V$ become higher order statistics of data (e.g. MSE in H$_V$ includes HOS of data).

- Members of H$_V$ are deterministic quantities even when x is r.v.
- Euclidean distance and QMI

$$D_{ED}(f,g) = \left\| V(f,.) - V(g,.)\right\|^2 \qquad QMI_{ED}(f_{1,2}, f_1 f_2) = \left\| V(f_{1,2},.) - V(f_1 f_2,.)\right\|^2$$

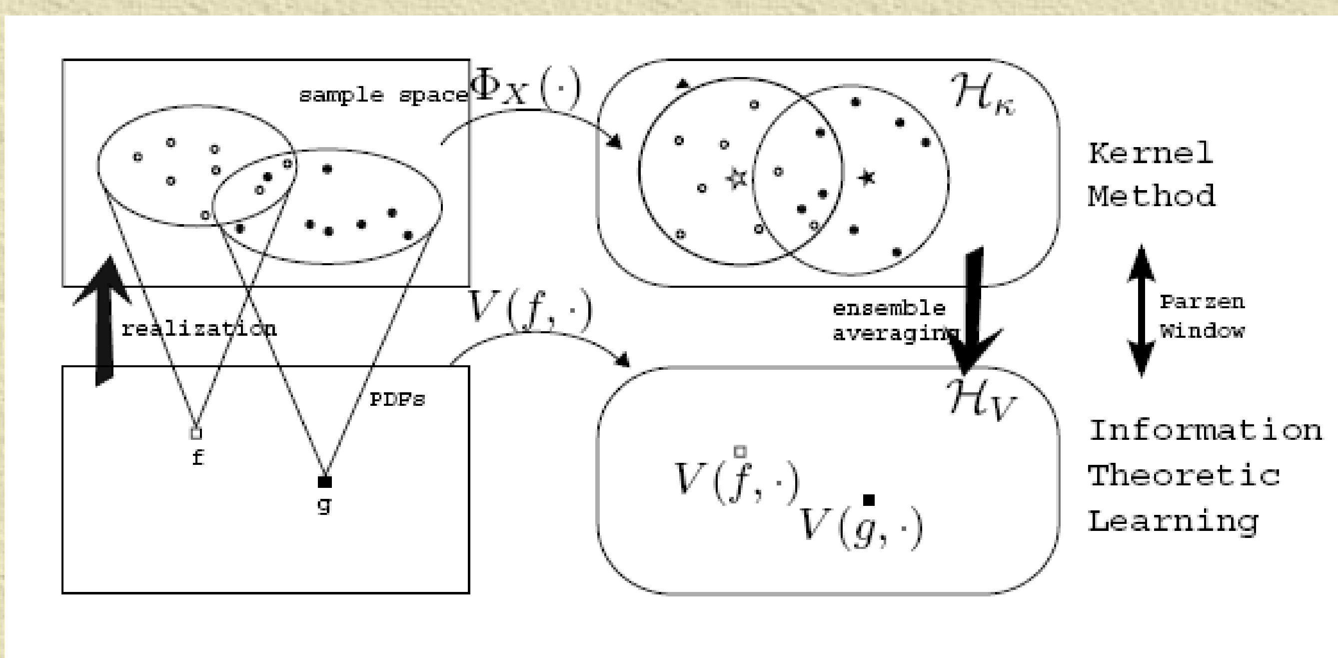
- Cauchy-Schwarz divergence and QMI

$$D(f,g) = -\log \frac{\left\langle V(f,.), V(g,.)\right\rangle_{H_V}}{\left\| V(f,.)\right\| . \left\| V(g,.)\right\|} = -\log(\cos\theta) \qquad QMI_{CS}(f_{1,2}, f_1 f_2) = -\log \frac{\left\langle V(f_{1,2},.), V(f_1 f_2,.)\right\rangle_{H_V}}{\left\| V(f_{1,2},.)\right\| . \left\| V(f_1 f_2,.)\right\|}$$

A RKHS for ITL

Relation between ITL and Kernel Methods thru H_V

There is a very tight relationship between H_V and H_K: By ensemble averaging of H_K we get **estimators** for the H_V statistical quantities. Therefore statistics in kernel space can be computed by ITL operators.



Correntropy:
A new generalized similarity measure

Correlation is one of the most widely used functions in signal processing.

But, correlation only quantifies similarity fully if the random variables are Gaussian distributed.

Can we define a new function that measures similarity but it is not restricted to second order statistics?

Use the kernel framework

Correntropy:
A new generalized similarity measure

Define correntropy of a stationary random process $\{x_t\}$ as

$$V_x(t,s) = E(\kappa(x_t - x_s)) = \int \kappa(x_t - x_s)\, p(x)dx$$

The name correntropy comes from the fact that the average over the lags (or the dimensions) is the information potential (the argument of Renyi's entropy)

For strictly stationary and ergodic r. p.

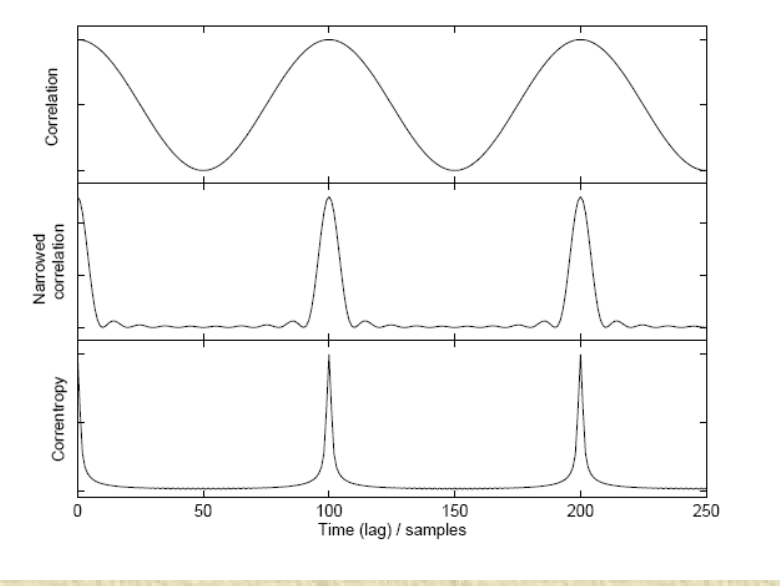
$$\hat{V}_m = \frac{1}{N}\sum_{n=1}^{N} \kappa(x_n - x_{n-m})$$

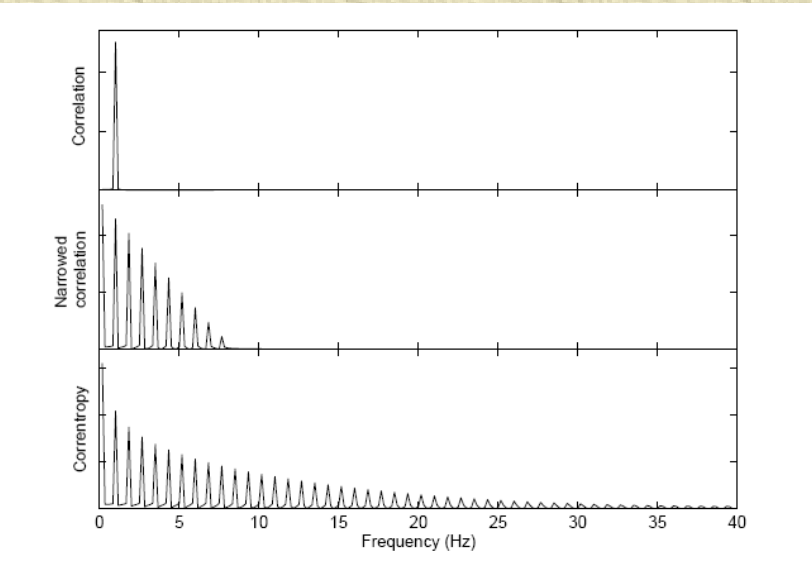
Correntropy can also be defined for pairs of random variables

Santamaria I., Pokharel P., Principe J., "Generalized Correlation Function: Definition, Properties and Application to Blind Equalization", IEEE Trans. Signal Proc. vol 54, no 6, pp 2187- 2186, 2006.

Correntropy:
A new generalized similarity measure

How does it look like? The sinewave





Correntropy:
A new generalized similarity measure

Properties of Correntropy:

- It has a maximum at the origin ($1/\sqrt{2\pi}\sigma$)
- It is a symmetric positive function
- Its mean value is the information potential
- Correntropy includes higher order moments of data

$$V(s,t) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n \sigma^{2n} n!} E\left\| x_s - x_t \right\|^{2n}$$

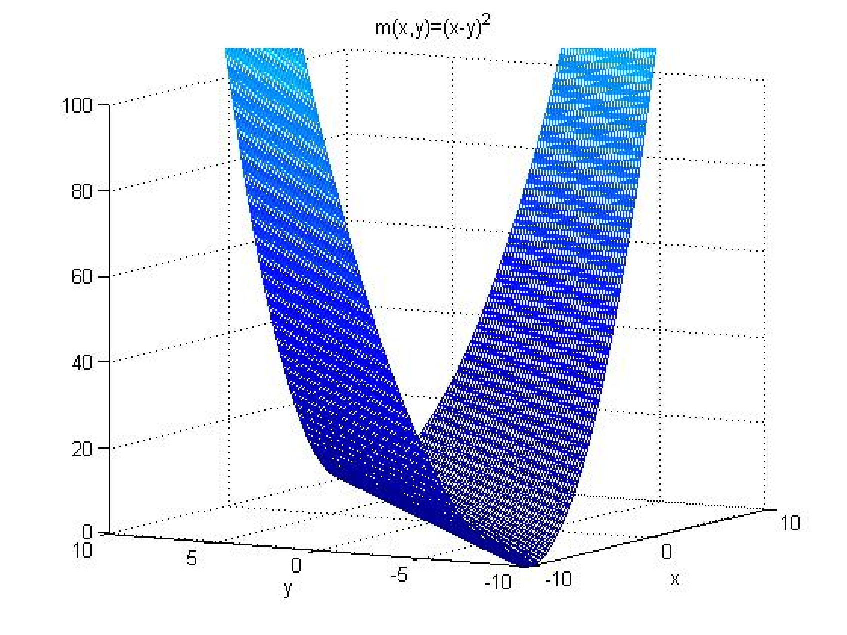
- The matrix whose elements are the correntopy at different lags is Toeplitz

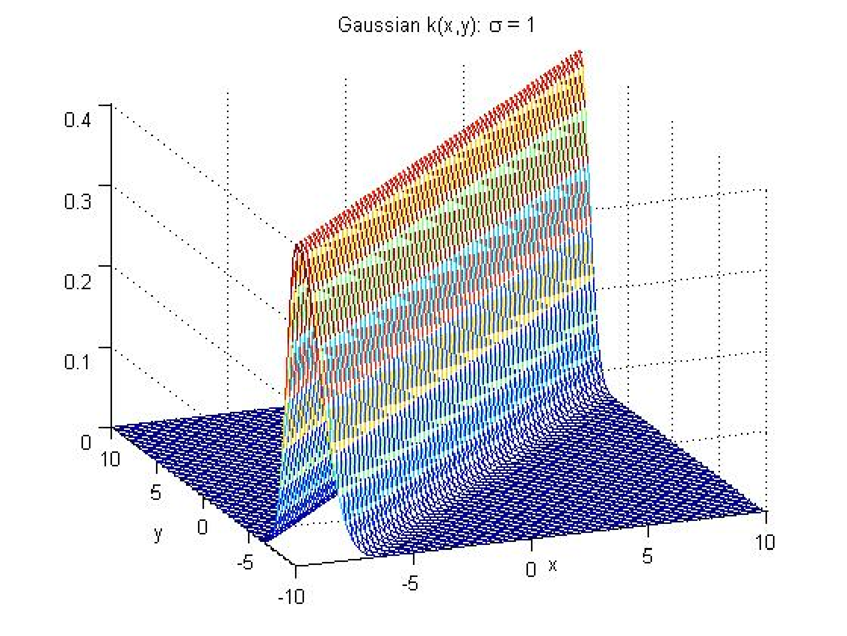
Correntropy:
A new generalized similarity measure

- Correntropy as a cost function versus MSE.

$$MSE(X,Y) = E[(X-Y)^2]$$
$$= \iint_{x,y} (x-y)^2 f_{XY}(x,y)dxdy$$
$$= \int_e e^2 f_E(e)de$$

$$V(X,Y) = E[k(X-Y)]$$
$$= \iint_{x,y} k(x-y) f_{XY}(x,y)dxdy$$
$$= \int_e k(e) f_E(e)de$$



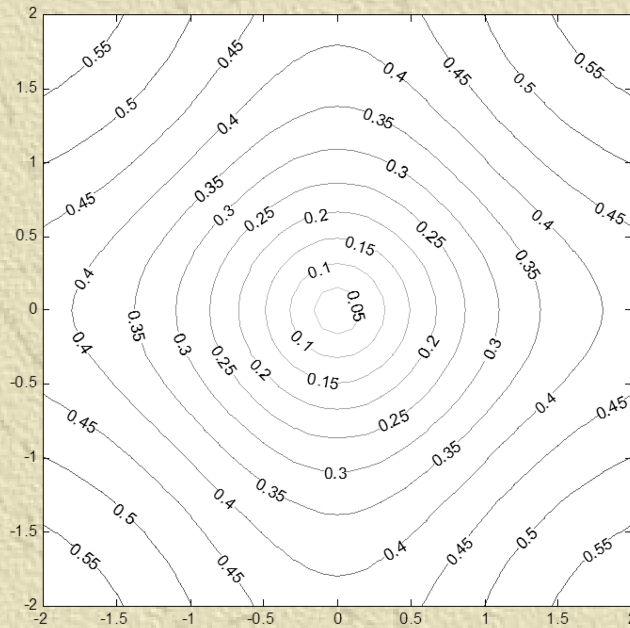


Correntropy:
A new generalized similarity measure

- Correntropy induces a metric (CIM) in the sample space defined by

$$CIM(X,Y) = (V(0,0) - V(X,Y))^{1/2}$$

- Therefore correntropy can be used as an alternative similarity criterion in the space of samples.



Liu W., Pokharel P., Principe J., "Correntropy: Properties and Applications in Non Gaussian Signal Processing", IEEE Trans. Sig. Proc, vol 55; # 11, pages 5286-5298

Correntropy:
A new generalized similarity measure

- Correntropy criterion implements M estimation of robust statistics. M estimation is a generalized maximum likelihood method.

$$\arg_\theta \min \sum_{i=1}^{N} \rho(x,\theta) \qquad \sum_{i=1}^{N} \psi(x_i, \hat{\theta}_M) = 0 \qquad \psi = \rho'$$

In adaptation the weighted square problem is defined as

$$\lim_{\theta} \sum_{i=1}^{N} w(e_i) e_i^2 \qquad w(e) = \rho'(e)/e$$

When

$$\rho(e) = (1 - \exp(-e^2/2\sigma^2))/\sqrt{2\pi}\sigma$$

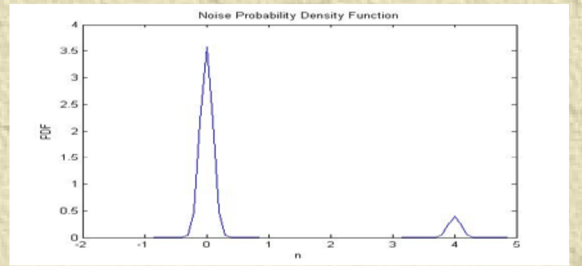
this leads to maximizing the correntropy of the error at the origin.

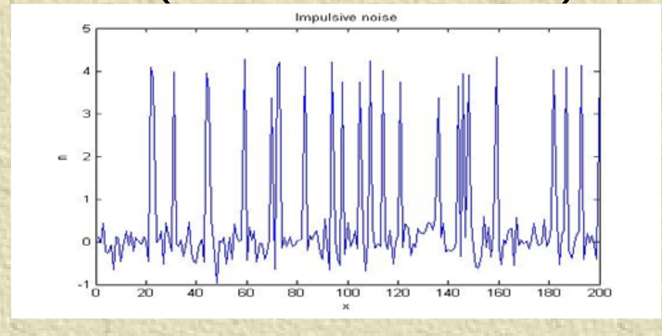
$$\min_\theta \sum_{i=1}^{N} \rho(e_i) = \min_\theta \sum_{i=1}^{N} (1 - \exp(-e_i^2/2\sigma^2))/\sqrt{2\pi}\sigma$$

$$\Leftrightarrow \max_\theta \sum_{i=1}^{N} \exp(-e_i^2/2\sigma^2)/\sqrt{2\pi}\sigma = \max_\theta \sum_{i=1}^{N} \kappa_\sigma(e_i)$$

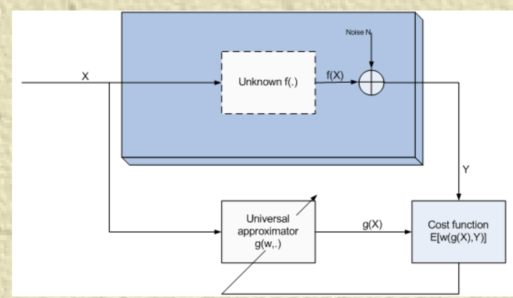
Correntropy:
A new generalized similarity measure

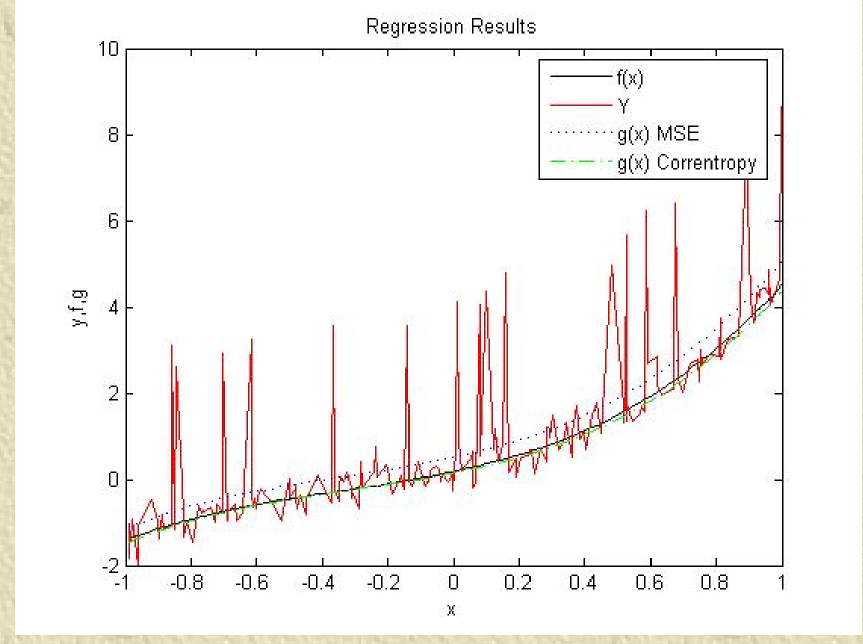
- Nonlinear regression with outliers (Middleton model)





- Polynomial approximator





Correntropy:
A new generalized similarity measure

Define **centered correntropy**

$$U(X,Y) = E_{X,Y}[\kappa(X-Y)] - E_X E_Y[\kappa(X-Y)] = \iint \kappa(x-y)\{dF_{X,Y}(x,y) - dF_X(x)dF_Y(y)\}$$

$$\hat{U}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x_i - y_i) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa(x_i - y_j)$$

Define **correntropy coefficient**

$$\eta(X,Y) = \frac{U(X,Y)}{\sqrt{U(X,X)U(Y,Y)}} \qquad \hat{\eta}(x,y) = \frac{\hat{U}(x,y)}{\sqrt{\hat{U}(x,x)\hat{U}(y,y)}}$$

Define **parametric correntropy** with $a,b \in R \quad a \neq 0$

$$V_{a,b}(X,Y) = E_{X,Y}[\kappa(aX+b-Y)] = \iint \kappa(ax+b-y)dF_{X,Y}(x,y)$$

Define **parametric centered correntropy**

$$U_{a,b}(X,Y) = E_{X,Y}[\kappa(aX+b-Y)] - E_X E_Y[\kappa(aX+b-Y)]$$

Define **Parametric Correntropy Coefficient**

$$\eta_{a,b}(X,Y) = \eta(aX+b,Y)$$

Rao M., Xu J., Seth S., Chen Y., Tagare M., Principe J., “Correntropy Dependence Measure”, Signal Processing 2010.

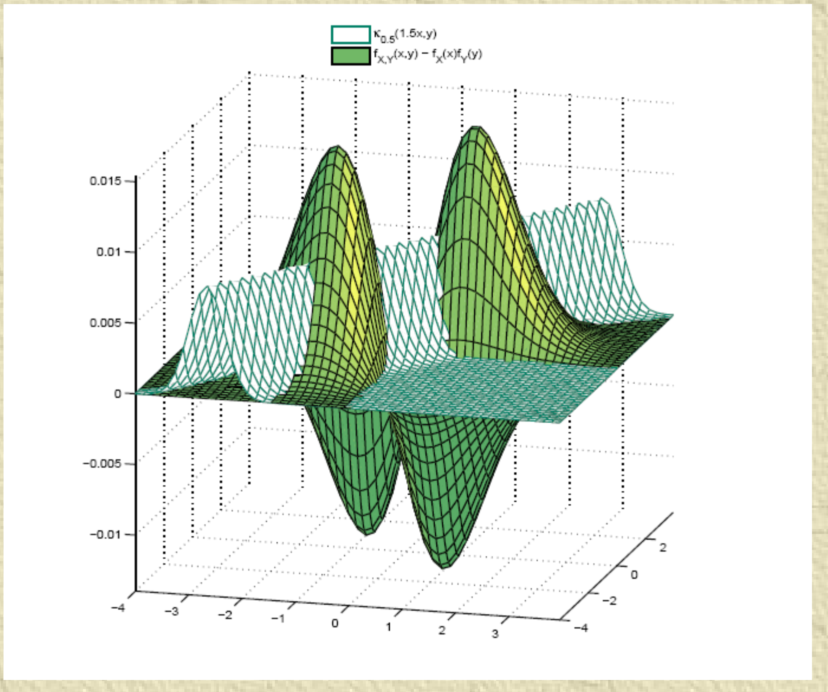
Correntropy:
Correntropy Dependence Measure

Theorem: Given two random variables X and Y: the parametric centered correntropy $U_{a,b}(X, Y) = 0$ for all *a*, *b* in R if and only if X and Y are independent.

Theorem: Given two random variables X and Y the parametric correntropy coefficient $\eta_{a,b}(X, Y) = 1$ for certain $a = a_0$ and $b = b_0$ if and only if $Y = a_0 X + b_0$.

Definition: Given two r.v. X and Y Correntropy Dependence Measure is defined as

$$\Gamma(X,Y) = \sup_{a,b \in R \quad a \neq 0} \left| \eta_{a,b}(X,Y) \right|$$

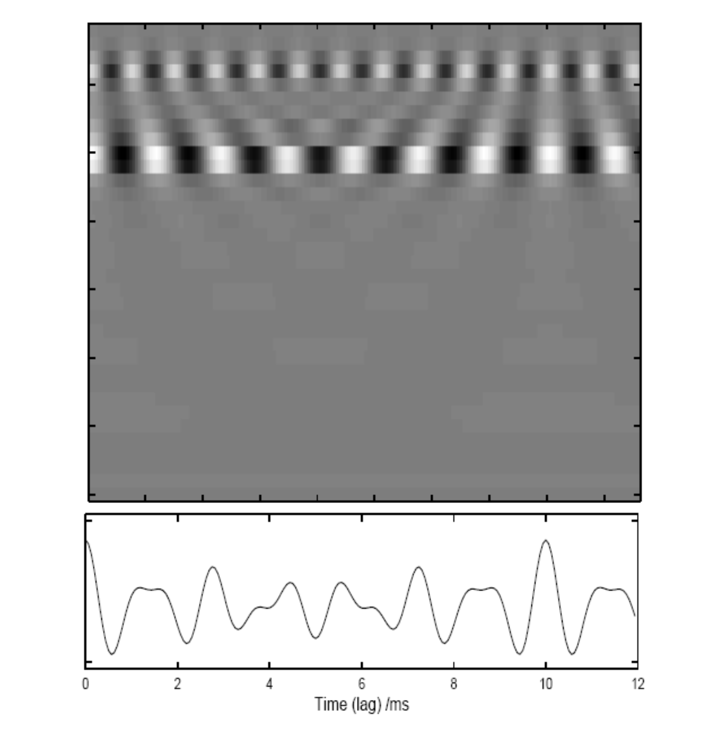


Applications of Correntropy
Correntropy based correlograms

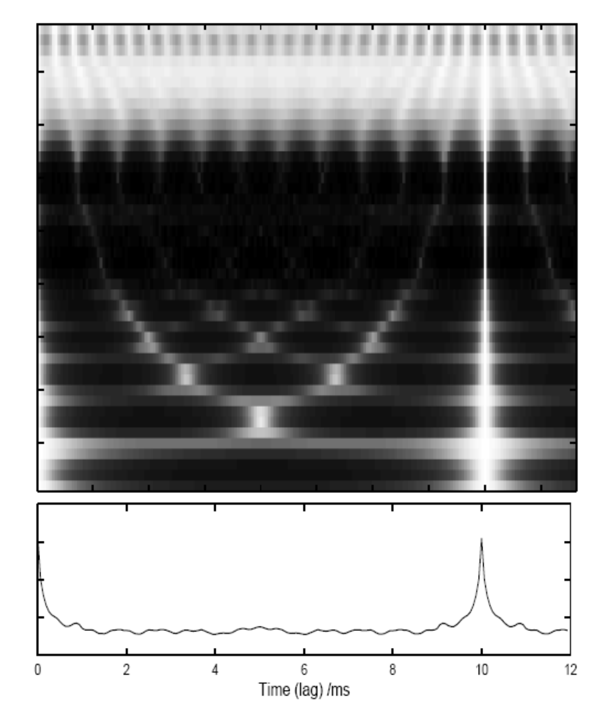
Correntropy can be used in computational auditory scene analysis (CASA), providing much better frequency resolution.

Figures show the correlogram from a 64 channel cochlea model for one (pitch=100Hz)).

Auto-correlation Function

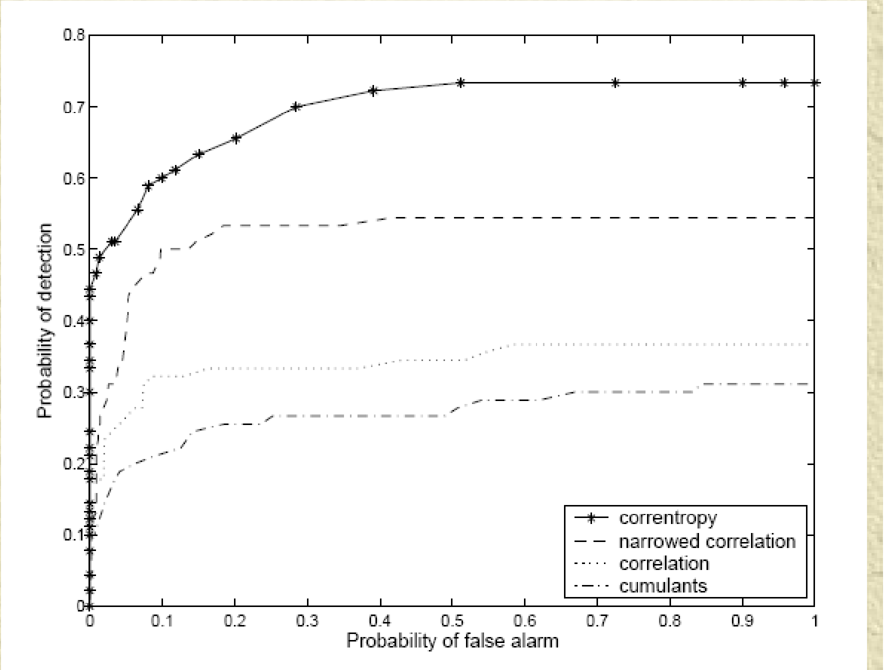


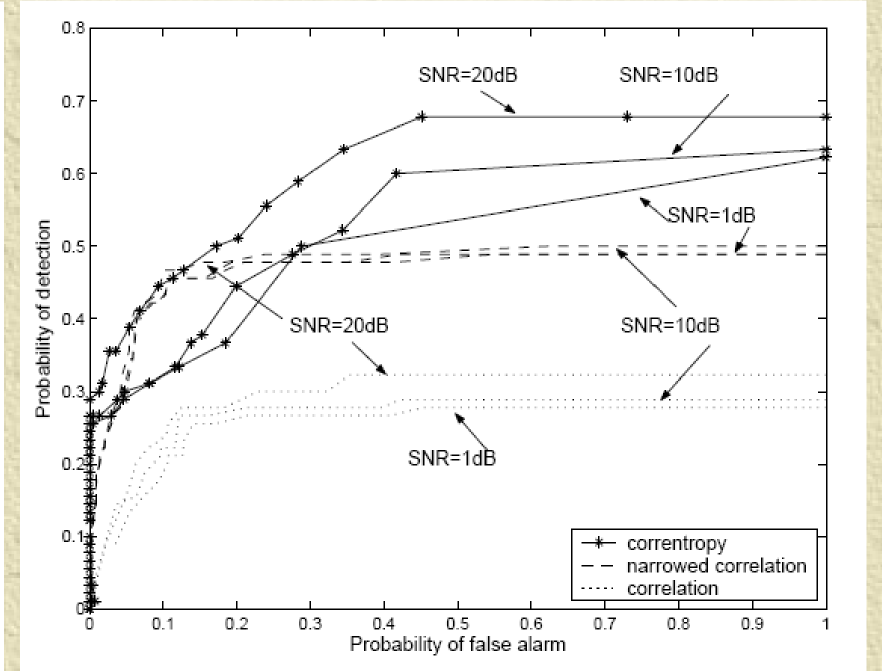
Auto-correntropy Function



Applications of Correntropy
Correntropy based correlograms

ROC for noiseless (L) and noisy (R) double vowel discrimiation





Applications of Correntropy
Matched Filtering

Matched filter computes the inner product between the received signal r(n) and the template s(n) ($R_{sr}(0)$).

The Correntropy MF computes

$$V_{rs}(0) = \frac{1}{N}\sum_{i=1}^{N} k(r_i - s_i)$$

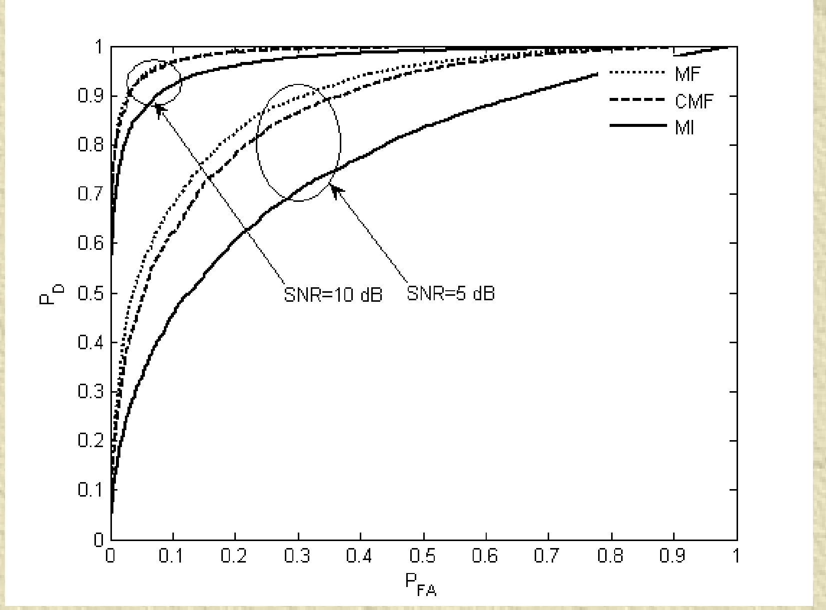
Hypothesis	received signal	Similarity value
H_0	$r_k = n_k$	$V_0 = \frac{1}{N\sqrt{2\pi\sigma^2}}\sum_{i=1}^{N} e^{-(s_i - n_i)^2/2\sigma^2}$
H_1	$r_k = s_k + n_k$	$V_1 = \frac{1}{N\sqrt{2\pi\sigma^2}}\sum_{i=1}^{N} e^{-n_i^2/2\sigma^2}$

(*Patent pending*)

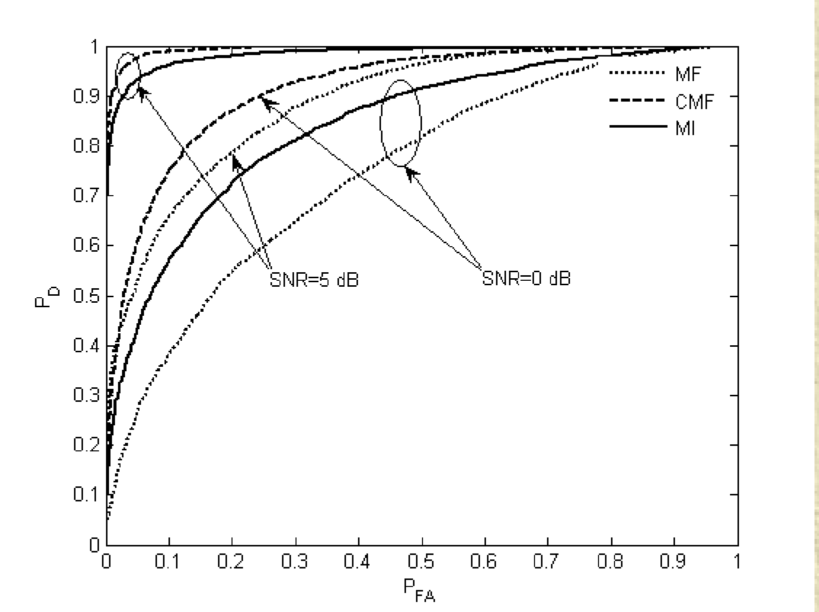
Applications of Correntropy
Matched Filtering

Linear Channels

White Gaussian noise



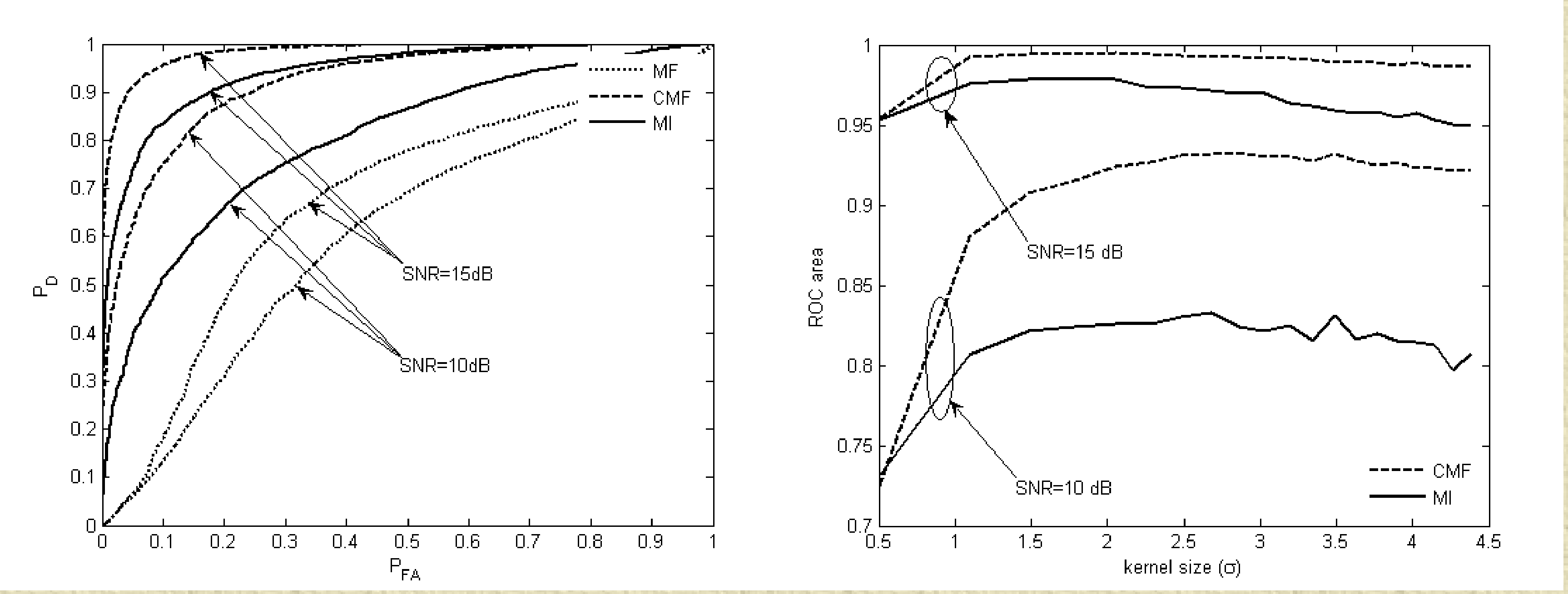
Impulsive noise



Template: binary sequence of length 20. kernel size using Silverman's rule.

Applications of Correntropy
Matched Filtering

Alpha stable noise (α=1.1), and the effect of kernel size



Template: binary sequence of length 20. kernel size using Silverman's rule.

Applications of Correntropy
Nonlinear temporal PCA

The Karhunen Loeve transform performs Principal Component Analysis (PCA) of the autocorrelation of the r. p.

$$X = \begin{bmatrix} x(1) & \dots & x(N) \\ \dots & \dots & \dots \\ x(L) & \dots & x(N+L-1) \end{bmatrix}_{LxN}$$

$$R = XX^T$$

$$\approx N \times \begin{bmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & r(0) & r(1) \\ r(L-1) & \cdots & r(1) & r(0) \end{bmatrix}_{L\times L}$$

$$K = X^T X$$

$$\approx L \times \begin{bmatrix} r(0) & r(1) & \cdots & r(N-1) \\ r(1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & r(0) & r(1) \\ r(N-1) & \cdots & r(1) & r(0) \end{bmatrix}_{N\times N}$$

$R = XX^T = UDD^T U^T$ $\qquad$ $K = X^T X = VD^T DV^T$ $\qquad$ D is LxN diagonal $\qquad$ $\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots \sqrt{\lambda_L}\}$

$$U_i^T X = \sqrt{\lambda_i} V_i^T, \quad i = 1, 2, \dots, L$$

KL can be also done by decomposing the Gram matrix K directly.

Applications of Correntropy
Nonlinear KL transform

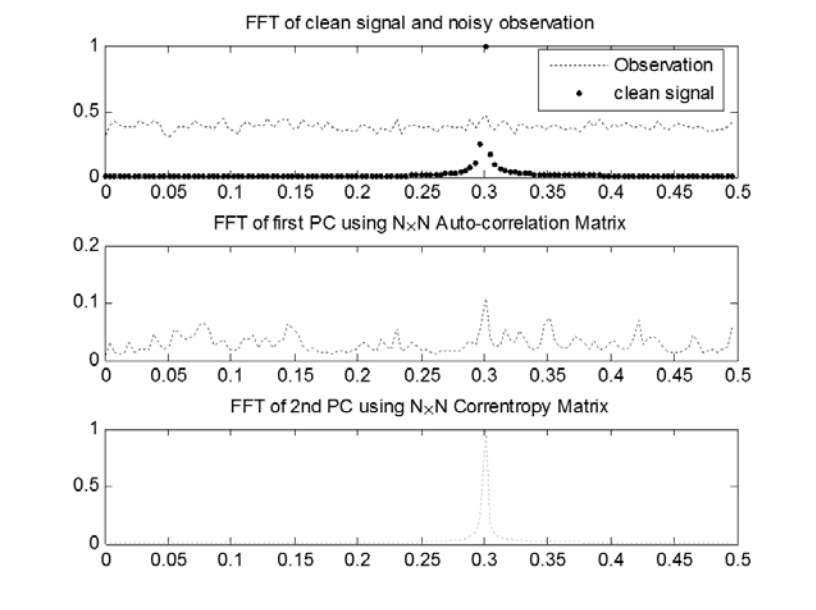
Since the autocorrelation function of the projected data in RKHS is given by correntropy, we can directly construct K with correntropy.

Example:

$$x(m) = A\sin(2\pi fm) + z(m)$$

where

$$p_N(n) = 0.8\times N(0,0.1) + 0.1\times N(4,0.1) + 0.1\times N(-4,0.1)$$



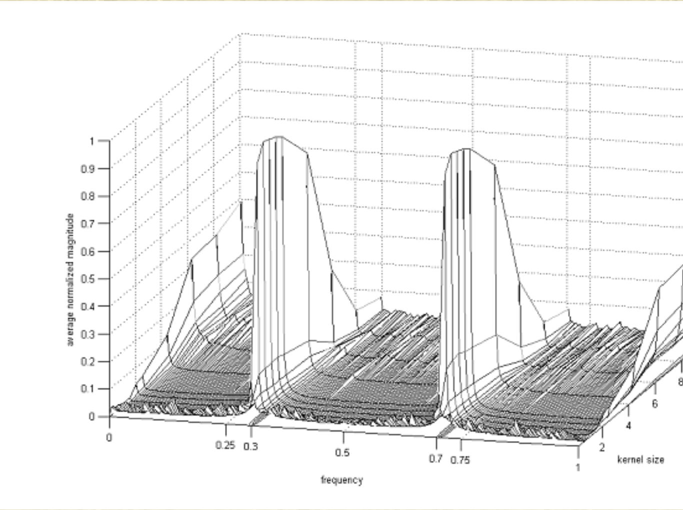
A	VPCA (2nd PC)	PCA by N-by-N (N=256)	PCA by L-by-L (L=4)	PCA by L-by-L (L=100)
0.2	100%	15%	3%	8%
0.25	100%	27%	6%	17%
0.5	100%	99%	47%	90%

1,000 Monte Carlo runs. σ=1

Applications of Correntropy
Correntropy Spectral Density

CSD is a function of the kernel size, and shows the difference between PSD (σ large) and the new spectral measure

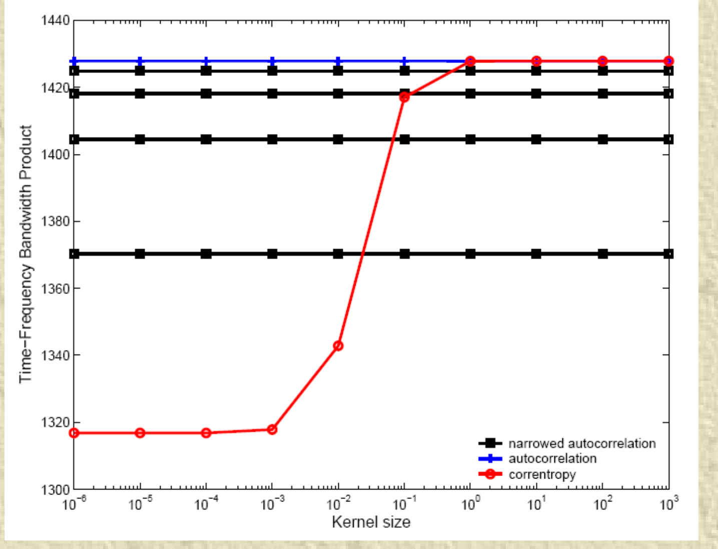
Average normalized amplitude



Kernel size

frequency

Time Bandwidth product



Kernel size

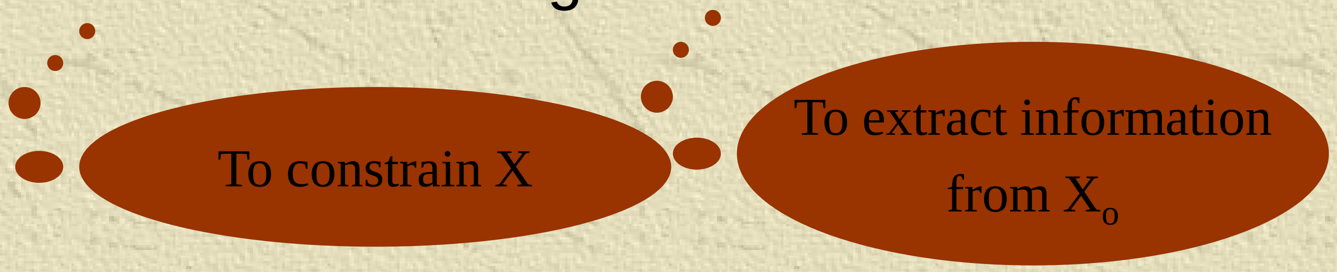
Principle of Relevant Information

STATEMENT

Consider a dataset $X_o = (x_o)_{i=1}^{N_o}$ with i.i.d. samples. We wish to find a new dataset $X = (x)_{i=1}^{N}$, $N \le N_o$ which captures predominant "structure" of the original dataset X_o .

INFORMATION THEORY FORMULATION

Cost Function = IT Goal +IT Regularization Term



Principle of Relevant Information

Principle of Relevant Information:

The conventional unsupervised learning algorithms for data representation (clustering, principal curves, vector quantization) are particular solutions to an information optimization problem that balances the minimization of data redundancy with the distortion between the original data and the solution, expressed by

$$\min_X L[p(x \mid x_0)] = H(X) + \lambda D_{KL}(X, X_0)$$

Principle of Relevant Information

- We will be using Renyi's quadratic entropy and its estimators to solve and apply in a nonparametric fashion the PRI.

$$J(X) = \min_X [H_2(X) + \lambda D_{CS}(X, X_0)] =$$
$$\min_X [(1-\lambda)H_2(X) + 2\lambda \log V(X, X_0) - \lambda H_2(X_0)]$$

- Drop last term because does not depend on *X*

$$J(X) = \min_X [(1-\lambda)H_2(X) - 2\lambda H(X; X_0)]$$

Case 1: λ=0

$$J(X) = \min_X \; H(X)$$

$H(X) = -\log\left(V(X)\right)$

$$J(X) = \max_X \; V(X)$$

Differentiating J(X) w.r.to $x_{k=\{1,2\ldots N\}}$

$$2\; F(x_k) = 0$$

$$x_k^{(\tau+1)} = m(x_k^{(\tau)}) = \frac{\sum_{j=1}^{N} G_\sigma\left(x_k - x_j\right)x_j}{\sum_{j=1}^{N} G_\sigma\left(x_k - x_j\right)}$$

GBMS

Case 2: λ=1

$$J(X) = \min_X \; H(X; X_o)$$

$$H(X;X_o) = -\log\left(V(X;X_o)\right)$$

$$J(X) = \max_X \; V(X; X_o)$$

Differentiating J(X) w.r.to $x_{k=\{1,2\ldots N\}}$

$$2\,F(x_k; X_o) = 0$$

$$x_k^{(\tau+1)} = m(x_k^{(\tau)}) = \frac{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right) x_{oj}}{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right)}$$

GMS

Case 3: Principal Curves

- Non linear extension of Principal Components (PCA)
- “Self-consistent” smooth curves which pass through the “middle” of a d-dimensional probability distribution or data cloud.
- Many definitions (Hestenes, etc…)

<u>A new definition</u> (Erdogmus et al.)

A point x is an element of the d-dimensional principal set ,denoted by ρ^d iff the transpose of the gradient $g(x)$ is orthonormal to at least (*n*-*d*) eigenvectors of the local Hessian $U(x)$ and $p(x)$ is a strict local maximum in the subspace spanned by these eigenvectors.

Case 3: PC continued

- ρ^0 is 0-dimensional principal set corresponding to modes of the data. ρ^1 is the 1-dimensional principal curve, ρ^2 is the 2-dimensional principal surface and so on….. $\rho^d \subset \rho^{d+1}$
-
- PRI satisfies this definition (experimentally).

$$J(X) = \min_X \; H(X) + \lambda D_{cs}(X, X_o)$$

Gives principal curves of 2D data for $1 < \lambda < 3$

Case 4: $\lambda \to \infty$

$$J(X) \to D_{cs}(X, X_o)$$

<u>Proof Outline</u>

- Start with equation $F(X) = \min_{X} D_{cs}(X, X_o)$
- Derive the fixed point update rule.
- Show that this is the same as taking $\lambda \to \infty$ in PRI fixed point

General case:

Cost Function

$$J(X) = \min_{X} \quad (1-\lambda)H(X) + 2\lambda H(X;X_o)$$

Rewriting gives

$$J(X) = \min_{X} \quad -(1-\lambda)\log\big(V(X)\big) - 2\lambda \log V(X;X_o)$$

Differentiating w.r.to $x_{k=\{1,2,\ldots,N\}}$

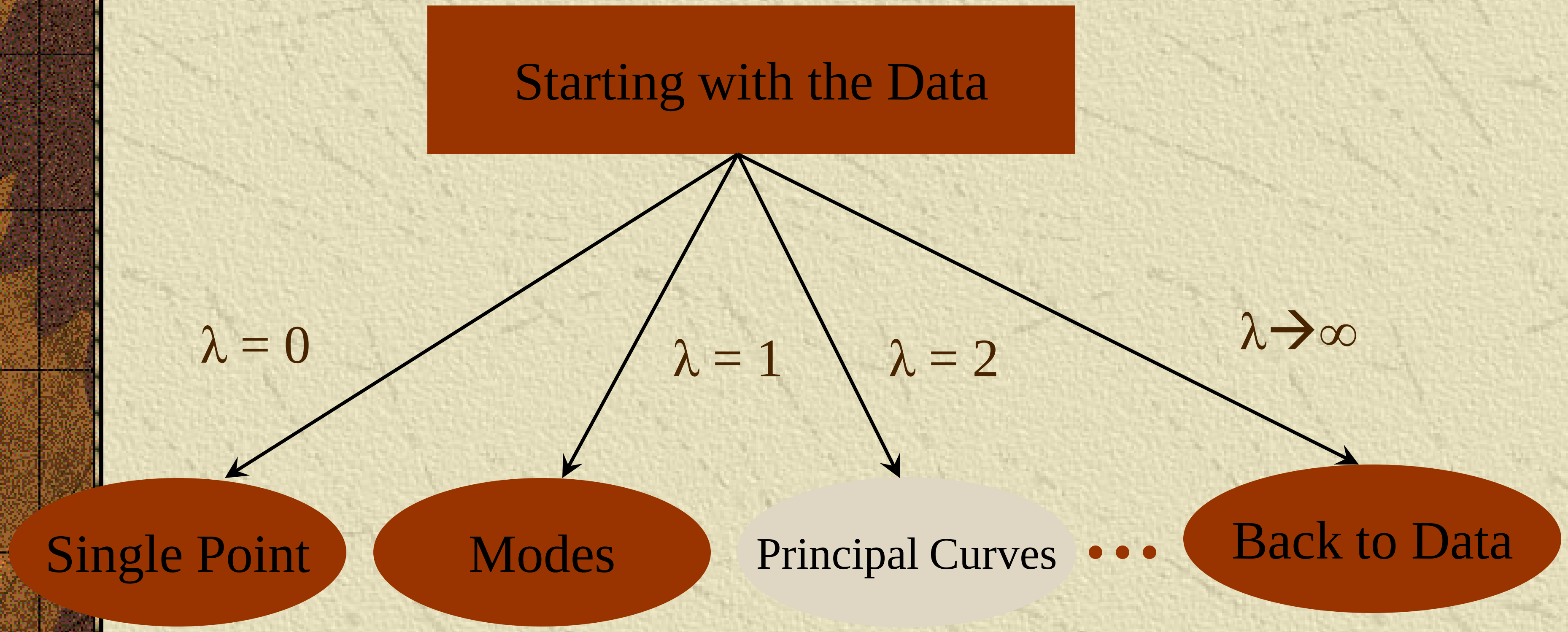
$$\frac{2(1-\lambda)}{V(X)} F(x_k) + \frac{2\lambda}{V(X;X_o)} F(x_k;X_o) = 0$$

PRI fixed point update

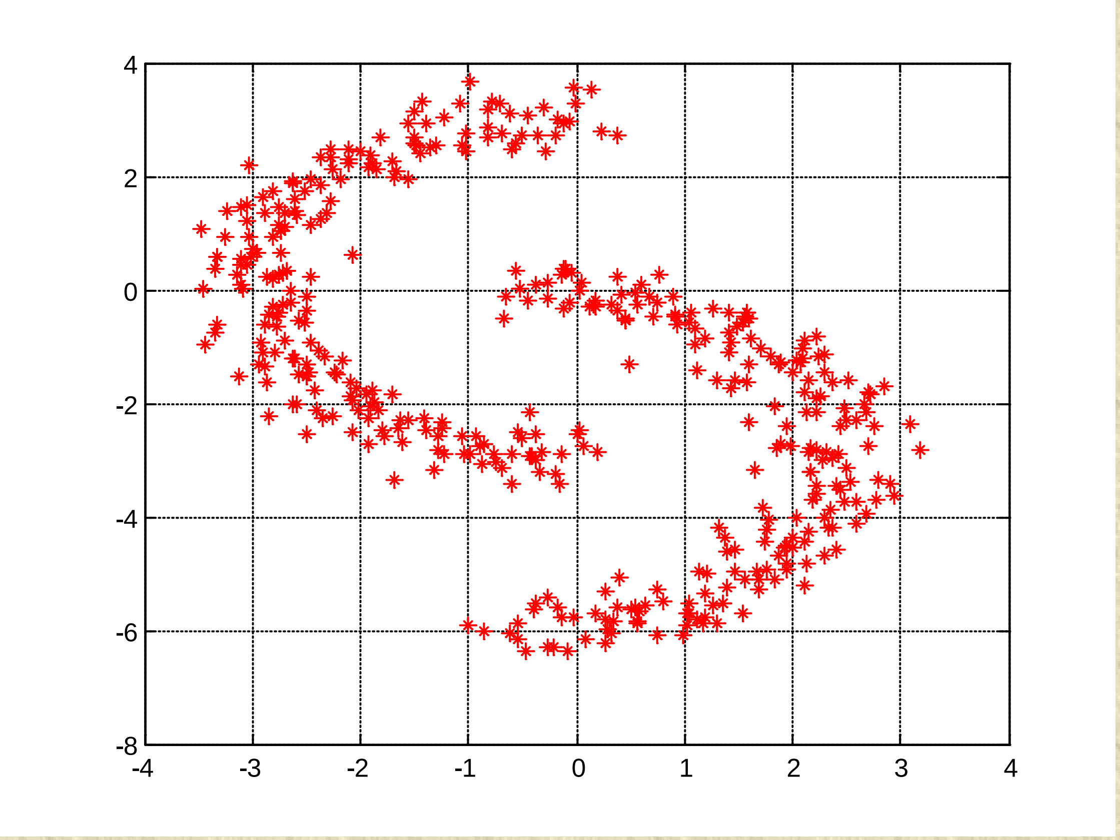
$$x_k^{(\tau+1)} = c\,\frac{(1-\lambda)}{\lambda}\,\frac{\sum_{j=1}^{N} G_\sigma\left(x_k - x_j\right)x_j}{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right)} + \frac{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right)x_{oj}}{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right)} - c\,\frac{(1-\lambda)}{\lambda}\,\frac{\sum_{j=1}^{N} G_\sigma\left(x_k - x_j\right)}{\sum_{j=1}^{N_o} G_\sigma\left(x_k - x_{oj}\right)}\,x_k$$

where $c = \frac{V(X;X_o)}{V(X)}\frac{N_o}{N}$

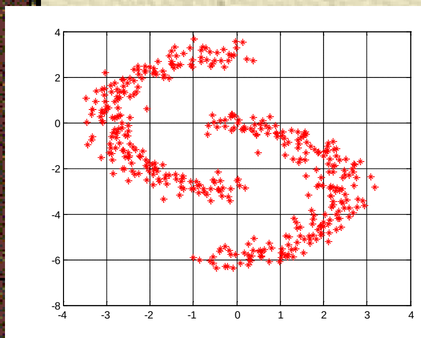
Summary

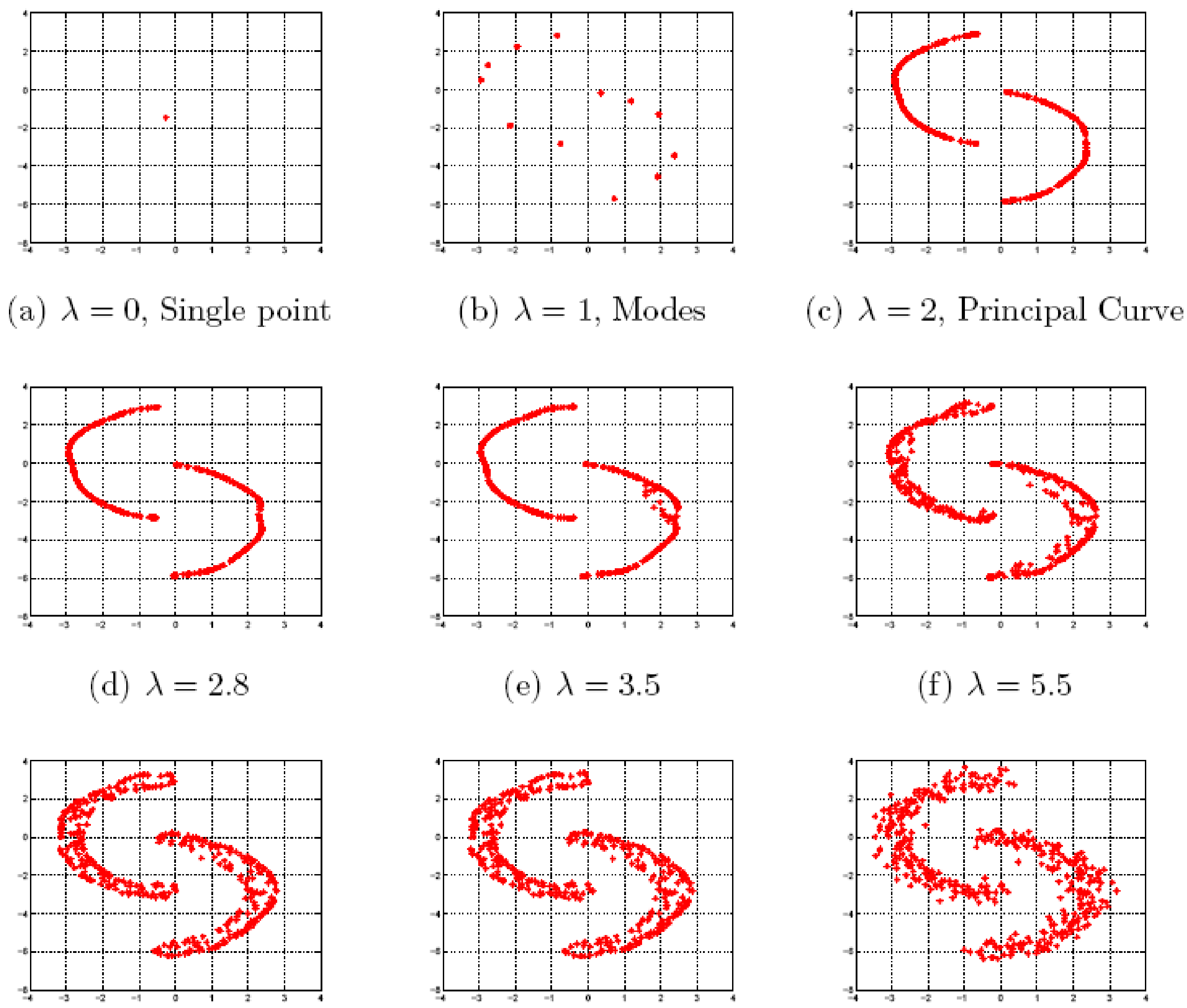

Starting with the Data
λ = 0
λ = 1
λ = 2
λ→∞
Single Point
Modes
Principal Curves
• • •
Back to Data

An Example



PRI result





(a) $\lambda = 0$, Single point

(b) $\lambda = 1$, Modes

(c) $\lambda = 2$, Principal Curve

(d) $\lambda = 2.8$

(e) $\lambda = 3.5$

(f) $\lambda = 5.5$

(g) $\lambda = 13.1$

(h) $\lambda = 20$

(i) $\lambda \to \infty$, The data

Conclusions

- Information Theoretic Learning took us beyond Gaussian statistics and MSE as cost functions.
  - ITL generalizes many of the statistical concepts we take for granted.
- Kernel methods implement shallow neural networks (RBFs) and extend easily the linear algorithms we all know.
  - KLMS is a simple algorithm for on-line learning of nonlinear systems
- Correntropy defines a new RKHS that seems to be very appropriate for nonlinear system identification and robust control
  - Correntropy may take us out of the local minimum of the (adaptive) design of optimum linear systems

For more information go to the website www.cnel.ufl.edu → ITL resource for tutorial, demos and downloadable MATLAB code