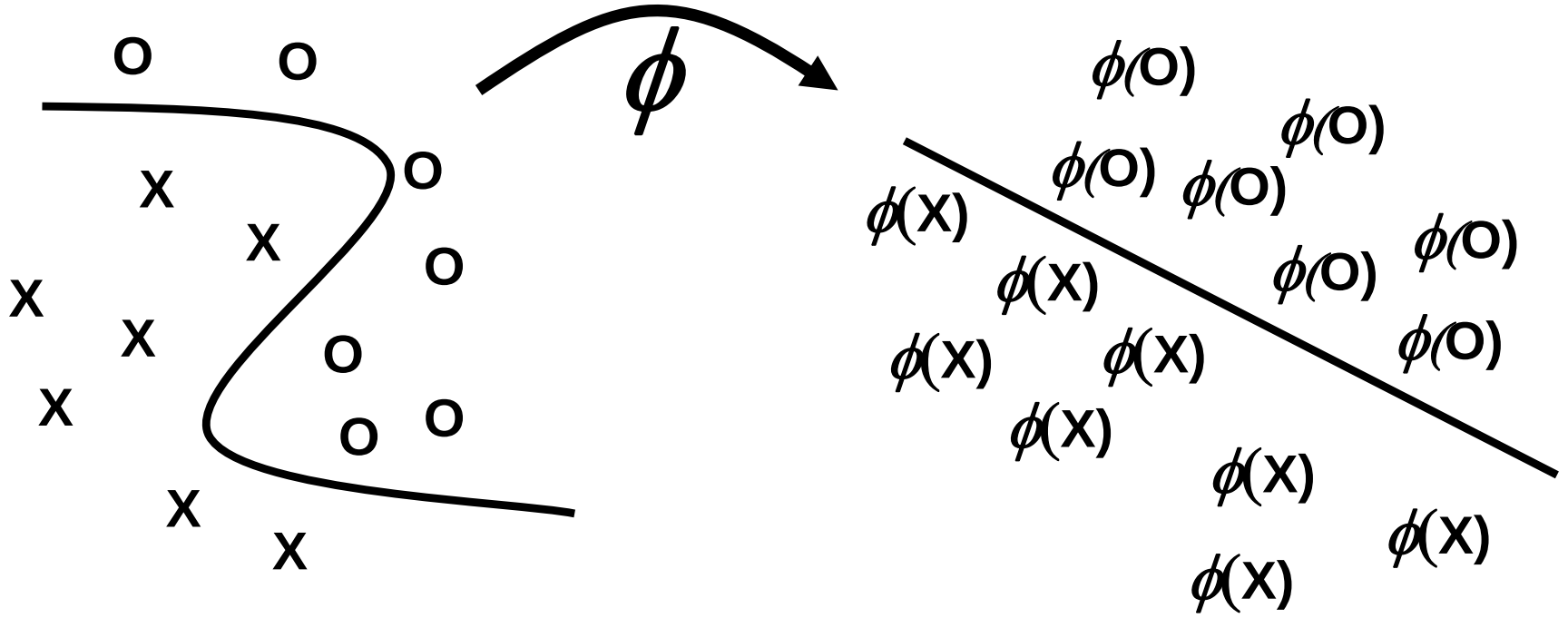
- Nonlinear problems (regression, classification, …) may be dealt with linearly by embedding the data in a higher-dimension space:



- A kernel is a function κ such that for all $\mathbf{x}, \mathbf{z} \in X$,

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

$$\mathbf{x} \rightarrow \phi(\mathbf{x}) \in F$$

F a vector space with an equipped with an inner product

- The "kernel trick" allows one to compute scalar products in a high or even infinite dimensional space with a limited number of computations.

- With $X = R^2$, $F = R^3$,

$$\phi : \mathbf{x} = (x_1, x_2) \rightarrow \phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$$< \phi(\mathbf{x}), \phi(\mathbf{z}) > = < (x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2) >$$

$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2$$

$$= (x_1x_1 + x_2z_2)^2 = < \mathbf{x}, \mathbf{z} >^2 = \kappa(\mathbf{x}, \mathbf{z})$$

Hence $\kappa(.,.)$ is a kernel function.

- With X consisting of all subsets of some set D, F = R, consider the kernel:

$$\kappa(A_1, A_2) = 2^{|A_1 \cap A_2|}$$

i.e. the number of common subsets A_1 and A_2 have.

- This kernel corresponds to a map to the vector space of dimension $2^{|D|}$ indexed by the subsets of D, with:

$$\phi_U(A) = \begin{cases} 1; & \text{if } U \subseteq A \\ 0; & \text{otherwise} \end{cases}$$

- For a finite set of input vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, all information on the mapping can be summarized in the Gram matrix $\mathbf{K}$ defined by:

$$\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$$

- If κ is used as a measure of similarity between vectors, the two extremes:

- only diagonal entries of $\mathbf{K}$ non zero
- All entries of $\mathbf{K}$ similar

are to be avoided.

- A function $\kappa : X \times X \rightarrow R$ either continuous or with a finite domain can be decomposed as

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

if and only if all Gram matrices formed with finite subsets of X are positive semi-definite.

- Note that if κ is a kernel, for all Gram matrices:

$$\mathbf{K} = \Phi\Phi^{\top} \text{ with } \Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_N)]^{\top}$$

And are thus positive semi-definite.

- For the "if" part, if κ satisfies the positive semi-definite property then it is possible to build a set of functions:

$$F = \left\{ \sum_{i=1}^{M} \alpha_i \kappa(\mathbf{x}_i, \bullet), M \text{ any integer} \right\}$$

- F is a vector space, and it can be equipped with an inner product. For:

$$f(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) \quad \text{and} \quad g(\mathbf{x}) = \sum_{j=1}^{L} \beta j \kappa(\mathbf{z}_j, \mathbf{x})$$

one defines:

$$< f, g > = \sum_{i=1}^{M} \sum_{j=1}^{L} \alpha_i \beta_j \kappa(\mathbf{x}_i, \mathbf{z}_j) = \sum_{i=1}^{M} \alpha_i g(\mathbf{x}_i) = \sum_{j=1}^{L} \beta_j f(\mathbf{z}_j)$$

If indeed all Gram matrices are positive semi-definite:

$$< f, g > = \sum_{i=1}^{M} \sum_{j=1}^{M} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\alpha}^{\mathsf{T}} \mathbf{K} \boldsymbol{\alpha} \geq 0$$

so $< \bullet , \bullet >$ is ineed an inner product.

- *Reproducing property* of the kernel:

$$< f, \kappa(\mathbf{x},\bullet) > = \sum_{i=1}^{M} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) = f(\mathbf{x})$$

- Mapping:

$$\phi : \mathbf{x} \in X \rightarrow \phi(\mathbf{x}) = \kappa(\mathbf{x},\bullet) \in F$$

- *Mercer's theorem*. For any valid kernel:

$$\kappa(\mathbf{x}, \mathbf{z}) = \sum_{k=1}^{\infty} \varphi_k(\mathbf{x}) \varphi_k(\mathbf{z})$$

- Kernel normalization:

$$\mathbf{x} \rightarrow \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}$$

$$\bar{\kappa}(\mathbf{x},\mathbf{z}) = \left\langle \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}, \frac{\phi(\mathbf{z})}{\|\phi(\mathbf{z})\|} \right\rangle = \frac{\kappa(\mathbf{x},\mathbf{z})}{\sqrt{\kappa(\mathbf{x},\mathbf{x})\kappa(\mathbf{z},\mathbf{z})}}$$

- Some recipes:

$$\kappa(\mathbf{x}, \mathbf{z}) = \kappa_1(\mathbf{x}, \mathbf{z}) + \kappa_2(\mathbf{x}, \mathbf{z})$$

$$\kappa(\mathbf{x}, \mathbf{z}) = a\, \kappa_1(\mathbf{x}, \mathbf{z}) \quad a > 0$$

$$\kappa(\mathbf{x}, \mathbf{z}) = \kappa_1(\mathbf{x}, \mathbf{z})\kappa_2(\mathbf{x}, \mathbf{z})$$

$$\kappa(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})\, f(\mathbf{z})$$

$$\kappa(\mathbf{x}, \mathbf{z}) = P(\kappa_1(\mathbf{x}, \mathbf{z})) \quad P \text{ polynomial with positive coefficients}$$

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp(\kappa_1(\mathbf{x}, \mathbf{z}))$$

- This kernel is expressed as (X must be a vector space with an inner product):

$$\kappa(\mathbf{x}, \mathbf{z}) = \exp(-||\mathbf{x} - \mathbf{z}||^2/(2\sigma^2))$$

Since $<\mathbf{x},\mathbf{z}>$ is a kernel, $\exp(<\mathbf{x},\mathbf{z}>/\sigma^2)$ is a kernel. If the latter is normalized:

$$\frac{\exp(<\mathbf{x},\mathbf{z}>/\sigma^2)}{\sqrt{\exp(||\mathbf{x}||^2/\sigma^2)\exp(||\mathbf{z}||^2/\sigma^2)}} =$$

$$\exp\left(\frac{<\mathbf{x},\mathbf{z}>}{\sigma^2} - \frac{<\mathbf{x},\mathbf{x}>}{2\sigma^2} - \frac{<\mathbf{z},\mathbf{z}>}{2\sigma^2}\right) = \exp\left(-\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}\right)$$

- Note that for this gaussian kernel the function from X to R:

$$f_{\mathbf{z}} : \mathbf{x} \rightarrow \kappa(\mathbf{x}, \mathbf{z})$$

is a radial basis function.

- The nonlinear mapping corresponding to that kernel maps X to an infinite-dimensional space due to:

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

- Classical probability density estimation presents several problems:

  - selection of the bins

  - non smooth nature

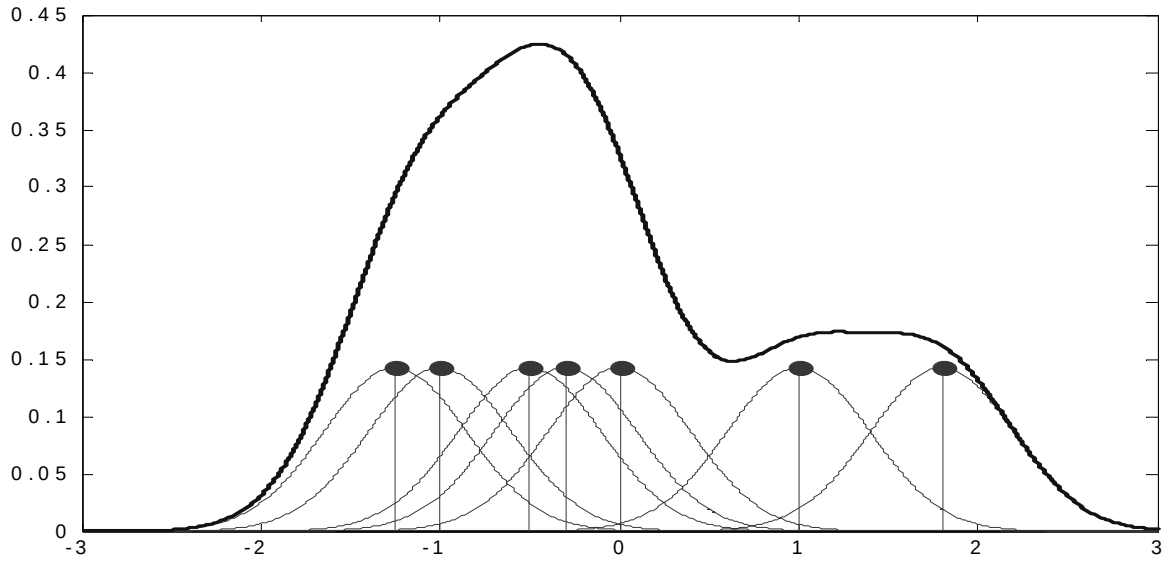
- In a univariate context, the kernel density estimate of a set of data points $\{x_1, x_2, \ldots, x_N\}$ is:

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{x - x_n}{h} \right)$$

where $K(.)$ is called the kernel function (no immediate connection with the previous slides to start with, but see later) and h is a bandwidth parameter defining the horizontal size of the scaled kernel function.

- Some popular kernels are the Epanechnikov, biweight, and triweight ones. But, do not be surprised, the most widely used is the Gaussian kernel.

- Of course the kernel must be normalized so that the integral of the estimate is one.
- Principle:



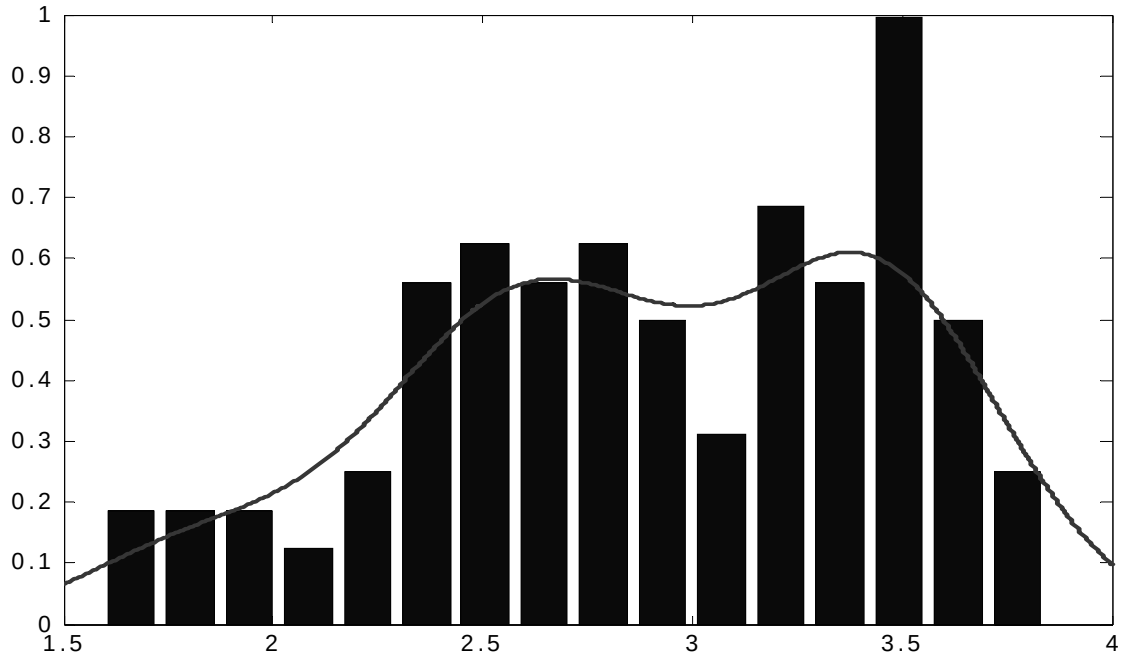
- In the one-dimensional case, an appropriate choice for the bandwidth parameter is:

$$h = 1.06sN^{-0.2}$$

With N the number of samples and s the sample standard deviation.

- In the multi-dimensional case, one can use a multi-dimensional Gaussian kernel with a bandwidth parameter as defined above for each coordinate.

- Example: log of the lynx time series



- The α-order Renyi's entropy for a probability density function $p(x)$ is defined as:

$$H_\alpha(p) = \frac{1}{1-\alpha}\log\int p^\alpha(x)dx = \frac{1}{1-\alpha}\log E[p^{\alpha-1}(x)]$$

- Taking the limit $\alpha \to 1$gives the Shannon entropy. The value $\alpha = 2$ gives:

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

- Using the Gaussian kernel pdf estimate with bandwidth *h:*

$$V_2(x) = \int p^2(x)dx = \int \left(\frac{1}{N} \sum_{n=1}^{N} \kappa_h (x - x_n)^2 \right) dx$$

$$= \frac{1}{N^2} \sum_{j=1}^{N} \sum_{i=1}^{N} \kappa_{h\sqrt{2}} (x_j - x_i)$$

- The integral of a product of Gaussians being a Gaussian. Note the change in the bandwitdh.

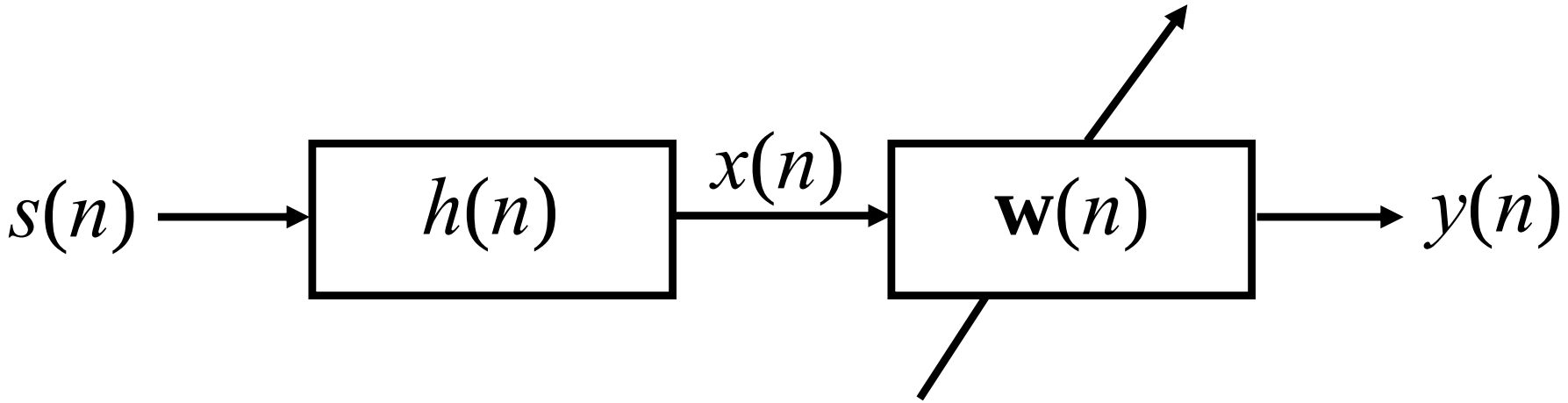
- Another approach consists in approximating the expected value by the sample mean in the definition of Renyi's entropy. One gets:

$$H_\alpha(p) = \frac{1}{1-\alpha}\log\left[\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N}\kappa_h(x_j - x_i\right)^{\alpha-1}\right]$$

- And for Shannon's entropy:

$$H_1(p) = -\frac{1}{N}\sum_{j=1}^{N}\log\left[\frac{1}{N}\sum_{i=1}^{N}\kappa_h(x_j - x_i)\right]$$

- The problem is illustrated as:



where the source $s(n)$ is a sequence of i.i.d. samples with unkown non-Gaussian pdf and the linear filter response is unknown.

- One tries to adapt $\mathbf{w}(n)$ so as to deconvolve $h(n)$, so that the pdf of $y(n)$ becomes as close as possible to that of $s(n)$.
- Principle: the pdf of the output $x(n)$ of the linear filter becomes closer to Gaussian due to the Central Limit effect. The Gaussian distribution has the largest entropy for a given variance. One tries to find $\mathbf{w}(n)$ that minimizes the entropy of $y(n)$.

- To have a scale-invariant criterion, one uses actually the criterion:

$$J(\mathbf{w}) = H_1(y) - \log[\mathrm{var}(y)]$$

- If a gradient scheme is used to minimize this criterion, the idea is to use an instantaneous estimate as in the LMS, so one replaces the entropy $\mathsf{E}[-\log f(y)]$ at time k by $-\log f(y_k)$.
- Of course, since the pdf $f(.)$ is unknown, one uses a kernel estimate.

- On a window of length L this estimate is:

$$\hat{f}(y_k) = \frac{1}{L} \sum_{i=k-L}^{k-1} \kappa_h(y_k - y_i)$$

- This leads to the entropy estimate gradient:

$$\frac{\partial \hat{H}_1(k)}{\partial \mathbf{w}} = - \frac{\sum_{i=k-L}^{k-1} \kappa_h'(y_k - y_i)\left(\frac{\partial y_k}{\partial \mathbf{w}} - \frac{\partial y_i}{\partial \mathbf{w}} \right)}{\sum_{i=k-L}^{k-1} \kappa_h(y_k - y_i)}$$

- Finally, since $y_k = \mathbf{w}^\mathsf{T}\mathbf{x}_k$:

$$\frac{\partial \hat{H}_1(k)}{\partial \mathbf{w}} = -\frac{\sum_{i=k-L}^{k-1} \kappa_h'(y_k - y_i)(\mathbf{x}_k - \mathbf{x}_i)}{\sum_{i=k-L}^{k-1} \kappa_h(y_k - y_i)}$$

- For L=1 and a Gaussian kernel one gets:

$$\frac{\partial \hat{H}_1(k)}{\partial \mathbf{w}} = -\frac{1}{h^2}(y_k - y_{k-1})(\mathbf{x}_k - \mathbf{x}_{k-1})$$

- With $\{\mathbf{x}_n\}$ a multivariate time series, it is possible to define through the kernel formulation a generalized correlation function (GCF). It is defined by:

$$V(p,q) = \mathsf{E}[\kappa_h(\mathbf{x}_p - \mathbf{x}_q)]$$

- Due to the nonlinearity induced by the kernel, $V(p,q)$ involves higher-order moments of the time series. When the kernel is Gaussian, all even-order moments are involved.

- Note that if $\kappa_h(\mathbf{x} - \mathbf{y})$ is a kernel, then:

$$\kappa_h(\mathbf{x}_p - \mathbf{x}_q) = \langle \phi(\mathbf{x}_p), \phi(\mathbf{x}_q) \rangle$$

so $\kappa_h(\mathbf{x}_p - \mathbf{x}_q)$ compares the two vectors by computing the inner product of their two images by ϕ.

- If additionally this kernel is normalized, then $\kappa_h(\mathbf{x}_p - \mathbf{x}_q)$ corresponds to the *cosine of the angle* between those two images.

- If the time series is strictly stationary, the GCF becomes a function of the time difference only:

$$V(m) = \mathsf{E}[\kappa_h(\mathbf{x}_n - \mathbf{x}_{n-m})]$$

- In practice, with only a finite sample set $\{\mathbf{x}_n\}$, $n = 1, \ldots, N$, one can obtain the estimate:

$$\hat{V}(m) = \frac{1}{N-m} \sum_{n=m+1}^{N} \kappa_h(\mathbf{x}_n - \mathbf{x}_{n-m})$$

- The Toeplitz matrix:

$$\mathbf{V}=\begin{bmatrix} V(0) & V(1) & \cdots & V(p-1) \\ V(1) & V(0) & \cdots & V(p-2) \\ \vdots & \ddots & \ddots & \vdots \\ V(p-1) & V(p-2) & \cdots & V(0) \end{bmatrix}$$

is positive definite as a sum of Gram matrices. This makes it possible to define a generalized power spectral density:

$$P(f)=\sum_{k=-\infty}^{\infty} V(k)\exp(-\mathrm{j}2\pi fk)$$

REFERENCES

- J. Shave-Taylor, *Kernel Methods for Pattern Analysis*, Cambridge U.P. 2004.
- D. Erdogmus and J.C. Principe, ″From linear adaptive filtering to nonlinear information processing,″ *IEEE Signal Processing Mag.*, vol. 23, no. 6, Nov. 2006, pp. 14-33.
- D. Erdogmus, K.E. Hild, and J.C. Principe, ″Online entropy manipulation: stochastic information gradient,″ *IEEE Signal Processing Lett.*, vol. 10, no. 8, Aug. 2003, pp. 242-245.
- I. santamaria, P.P. Pokharel, and J.C. Principe, ″Generalized correlation function: definition, properties, and application to blind equalization,″ *IEEE Trans. Signal Processing*, vol. 54, no. 6, June 2006, pp. 2187-2197.
- S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller, ″Kernel-based nonlinear blind source separation,″ *Neural Computation*, vol. 15, 2003, pp. 1089-1124.