Marc G. Genton (2001)

Classes of Kernels for Machine Learning: A Statistics Perspective

Randy J

8/15/2021

Introduction

Kernels allow to map the data into a high dimensional feature space in order to increase the computational power of linear machine.

Example algorithms

- Support vector machines
- Kernel principal component analysis
- Kernel Gram-Schmidt
- Bayes point machines
- Gaussian processes

Kernel K

$\forall\ \boldsymbol{x},\ \boldsymbol{z} \in \boldsymbol{X} \subset \Re^d :\ K(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\Phi}(\boldsymbol{x}),\ \boldsymbol{\Phi}(\boldsymbol{z}) \rangle$, where Φ is a nonlinear (or sometimes linear) map from the input space $\boldsymbol{X}$ to the feature space $\mathfrak{F}$, and $\langle .,\ . \rangle$ is an inner product.

- Kernel must be symmetric: $K(\boldsymbol{x},\ \boldsymbol{z}) = K(\boldsymbol{z},\ \boldsymbol{x})$
- Kernel also satisfies the Cauchy-Schwartz inequality: $K^2(\boldsymbol{x},\ \boldsymbol{z}) \leq K(\boldsymbol{x},\ \boldsymbol{x}) K(\boldsymbol{z},\ \boldsymbol{z})$.
- *Mercer (1909)*: a necessary and sufficient condition for a symmetric function $K(\boldsymbol{x},\ \boldsymbol{z})$ to be a kernel is that it be positive definite. $\forall\ x_1,\ ...,\ x_l$ and real numbers $\lambda_1,\ ...,\ \lambda_l$, the function K must satisfy: $\sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 0 \quad (1)$

Symmetric positive definite functions are called covariances in statistics.

Positive defintie kernel properties

1. If K_1, K_2 are two kernels, and a_1, a_2 are two positive real numbers, then: $K(\boldsymbol{x},\ \boldsymbol{z}) = a_1 K_1(\boldsymbol{x},\ \boldsymbol{z}) + a_2 K_2(\boldsymbol{x},\ \boldsymbol{z}) \quad (2)$ is a kernel.
2. The multiplication of two kernels K_1 and K_2 yields a kernel: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_1(\boldsymbol{x},\ \boldsymbol{z}) K_2(\boldsymbol{x},\ \boldsymbol{z}) \quad (3)$
3. Properties (2) and (3) imply that any polynomial with positive coefficients, $pol^+(x) = \{\sum_{i=1}^{n} \alpha_i x^i | n \in \mathfrak{N},\ \alpha_1,\ ...,\ \alpha_n \in \mathfrak{R}^+\}$, evaluatedat a kernel K_1, yields a kernel: $K(\boldsymbol{x},\ \boldsymbol{z}) = pol^+(K_1(\boldsymbol{x},\ \boldsymbol{z})) \quad (4)$

4. In particular, we have that: $K(\boldsymbol{x},\ \boldsymbol{z}) = \exp(K_1(\boldsymbol{x}, \boldsymbol{z}))$ (5) is a kernel by taking the limit of the series expansion of the exponential function
5. If g is a real-valued function on $\boldsymbol{X}$, then $K(\boldsymbol{x},\ \boldsymbol{z}) = g(\boldsymbol{x})g(\boldsymbol{z})$ (6) is a kernel.
6. If ψ is an $\mathfrak{R}_p$-valued function on $\boldsymbol{X}$ and K_3 is a kernel on $\mathfrak{R}_p \times \mathfrak{R}_p$, then: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_3(\psi(\boldsymbol{x}), \psi(\boldsymbol{z}))$ (7) is also a kernel
7. If $\boldsymbol{A}$ is a positive definite matrix of size $d \times d$, then: $K(\boldsymbol{x},\ \boldsymbol{z}) = \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{z}$ (8) is a kernel

Extra property to construct kernels

- Let h be a real-valued function on $\boldsymbol{X}$, positive, with minimum at 0 (that is, h is a variance function). Then:
  $K(\boldsymbol{x},\ \boldsymbol{z}) = \frac{h(\boldsymbol{x}+\boldsymbol{z})-h(\boldsymbol{x}-\boldsymbol{z})}{4}$ (9) is a kernel.
- The justification of (9) comes from the following identity for two random variables Y_1 and Y_2: $Cov[Y_1,\ Y_2] = \frac{Var[Y_1+Y_2]-Var[Y_1-Y_2]}{4}$.
- For instance, $h(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{x}$. From (9), we obtain the kernel:
  $K(\boldsymbol{x},\ \boldsymbol{z}) = \frac{(\boldsymbol{x}+\boldsymbol{z})^\top(\boldsymbol{x}+\boldsymbol{z})-(\boldsymbol{x}-\boldsymbol{z})^\top(\boldsymbol{x}-\boldsymbol{z})}{4} = \boldsymbol{x}^\top\boldsymbol{z}$.

Spectral distribution and density functions

Fourier transformation

Given a time series $\{x_t\}$, its Fourier transformation is:

$$x(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} e^{-it\omega} x(t)$$

and the inverse Fourier transform is:

$$x(t) = \int_{-\pi}^{\pi} e^{it\omega} x(\omega) d\omega$$

- ω denote the frequency $(-\pi < \omega < \pi)$
- T denote the period: the minimum time that it takes the wave to go through a whole cycle
- $T = 2\pi/\omega$
- Given any integer number z, $x(t) = x(t + zT)$
- ϕ denote the phase: the amount that a wave is shifted.

Wiener-Khintchine Theorem

- The *Wiener-Khintchine Theorem* says that if $\gamma(k)$ is the autocovariance function of X_t, there must exist a monotonically increasing function $F(\omega)$ such that

$$\gamma(k) = \int_0^{\pi} \cos(\omega_k) dF(\omega)$$

- $F(\omega)$ is the *spectral distribution function* of the process X_t.
- For $k = 0$, $\gamma(0) = \int_0^{\pi} dF(\omega) = F(\pi) = \sigma_x^2$ so all other variation in the process is for $0 < \omega < \pi$.
- We can redefine the spectral distribution function as: $F^*(\omega) = \frac{F(\omega)}{\sigma_x^2}$
- $F^*(\omega)$ is the proportion of variance accounted by ω.

CDF and PSF

- $F^*(0) = 0$, $F^*(\pi) = 1$ and since $F(\omega)$ is monotonically increasing then $F^*(\omega)$ is a cummulative distribution function (CDF).
- The *spectral density function* is denoted by $f(\omega)$ and defined as:

$$f(\omega) = \frac{dF(\omega)}{d\omega};\ 0 < \omega < \pi$$

- This function is also known as the *power spectral function* or *spectrum*
- The existence of $f(\omega)$ is under the assumption that the spectral distribution function is differentiable everywhere (except in a set of measure zero).
- *Wold's Theorem*: an alternative representation for the covariance function $\gamma(k) = \int_0^\pi \cos(\omega_k) f(\omega) d\omega$.

Kernel and Spectrum

- If the spectrum has a "peak" at ω_0, this implies that ω_0 is an important frequency of the process X_t.
- The spectrum or spectral density is a theoretical function of the process X_t. In practice, the spectrum is usually unknown and we use the *periodogram* to estimate it.
- There is an inverse relationship between the $f(\omega)$ and $\gamma(k)$, $f(\omega) = \frac{1}{\pi} \sum_{k=-\infty}^{\infty} \gamma(k) e^{-i\omega k}$, so the spectrum is the Fourier transformation of the autocovariance function.

Fourier transformation

- From complex analysis, recall that $e^{-i\omega k} = \cos(\omega k) - \sin(\omega k)i$
- This implies that

$$f(\omega) = \frac{1}{\pi}[\gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k)\cos(\omega k)]$$

- The normalized spectral density $f^*(\omega)$ is defined as: $f^*(\omega) = \frac{f(\omega)}{\sigma_x^2} = \frac{dF^*(\omega)}{d\omega}$
- $f^*(\omega) = \frac{1}{\pi}[1 + 2\sum_{k=1}^{\infty}\rho(k)\cos(\omega k)]$ so the normalized spectrum is the Fourier transform of the autocorrelation function (ACF).

Stationary Kernels

- A stationary kernel is one which is translation invariant: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_s(\boldsymbol{x} - \boldsymbol{z})$, which only depends on the lag vector.
- Also referred as ansiotropic stationary kernel, to emphasize the dependence on both direction and length of the lag vector.
- *Bochner (1955)*: $K_s(\boldsymbol{x} - \boldsymbol{z})$ is positive definite in $\Re^d$ iff it has the form: $K_s(\boldsymbol{x} - \boldsymbol{z}) = \int_{\Re^d} \cos\left(\boldsymbol{\omega}^\top(\boldsymbol{x} - \boldsymbol{z})\right) F(d\boldsymbol{\omega}) \quad (10)$, where F is a positive finite measure.
- The quantity $F/K_S(\mathbf{0})$ is called the spectral distribution function. *The Fourier dual in GPML book*. Note that (10) is simply the Fourier transform of F.

Isotropic (or homogeneous) stationary kernel

- When a stationary kernel depends only on the norm of the lag vector between two examples, and not on the direction, then the kernel is said to be isotropic (or homogeneous), and is thus only a function of distance: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_I(\|x - z\|)$.
- The spectral representation of isotropic stationary kernels has been derived from *Bochner's theorem (Bochner, 1955)* by Yaglom (1957):

$$K_I(\|x - z\|) = \int_0^\infty \Omega_d(\omega\|x - z\|)F(d\omega) \qquad (11)$$

- where $\Omega_d(x) = \left(\frac{2}{x}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{(d-2)/2}(x)$, form a basis for functions in $\Re^d$. Here F is any nondecreasing bounded function, $\Gamma(d/2)$ is the gamma function, and J_v is the Bessel function of the first kind of order v.

Isotropic (or homogeneous) stationary kernel

- Some familiar examples of Ω_d are $\Omega_1(x) = \cos(x)$, $\Omega_2(x) = J_0(x)$, and $\Omega_3(x) = \sin(x)/x$.
- By choosing a nondecreasing bounded function F (or its derivative f), we can derive the corresponding kernel from (11). For instance in $\Re^1$, with the spectral density $f(\omega) = (1 - \cos(\omega))/(\pi\omega_2)$, we derive the triangular kernel:

$$K_I(\boldsymbol{x} - \boldsymbol{z}) = \int_0^\infty \cos(\omega|x - z|)\frac{1 - \cos(\omega)}{\pi\omega^2}d\omega = \frac{1}{2}(1 - |x - z|)^+$$

where $(x)^+ = max(x,\ 0)$ (see *Figure 1*).

Spectral density and kernel

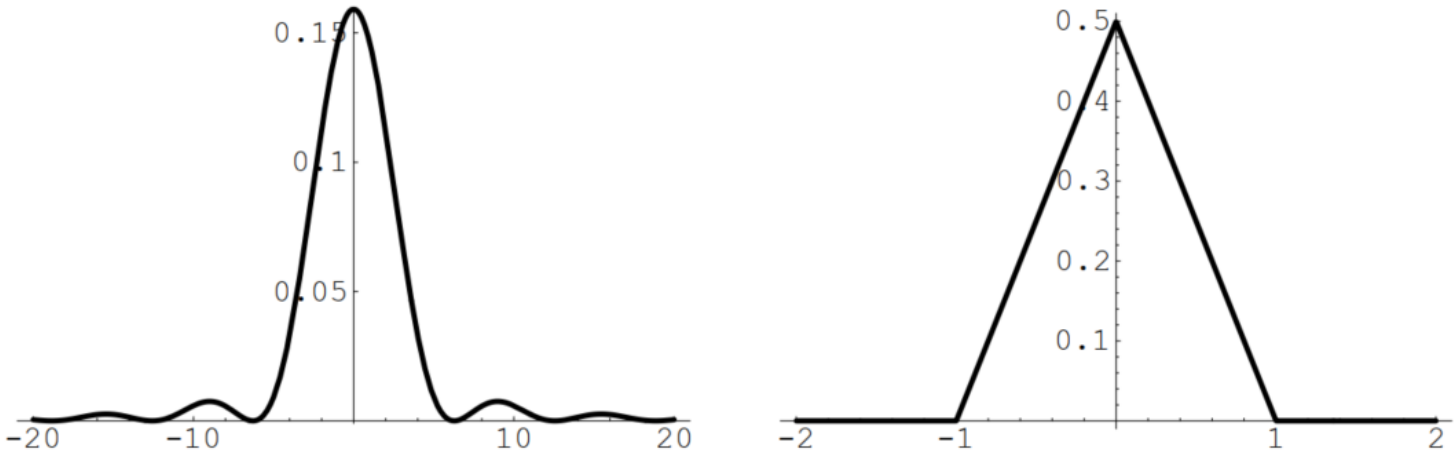


Figure 1: The spectral density $f(\omega) = (1 - \cos(\omega))/(\pi\omega^2)$ (left) and its corresponding isotropic stationary kernel $K_I(x - z) = (1 - |x - z|)^+/2$ (right).

- Note that an isotropic stationary kernel obtained with Ω_d is positive definite in $\Re^d$ and in lower dimensions, but not necessarily in higher dimensions.
- For example, the kernel $K_I(\boldsymbol{x} - \boldsymbol{z}) = (1 - |x - z|)^+/2$ is positive definite in $\Re^1$ but not in $\Re^2$.
- It is interesting to remark from (11) that an isotropic stationary kernel has a lower bound *(Stein, 1999)*: $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|)/K_I(0) \geq \inf_{x \geq 0} \Omega_d(x)$, thus yielding:
  - $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|)/K_I(0) \geq -1 \; in \; \Re^1$
  - $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|)/K_I(0) \geq -0.403 \; in \; \Re^2$
  - $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|)/K_I(0) \geq -0.218 \; in \; \Re^3$
  - $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|)/K_I(0) \geq 0 \; in \; \Re^\infty$

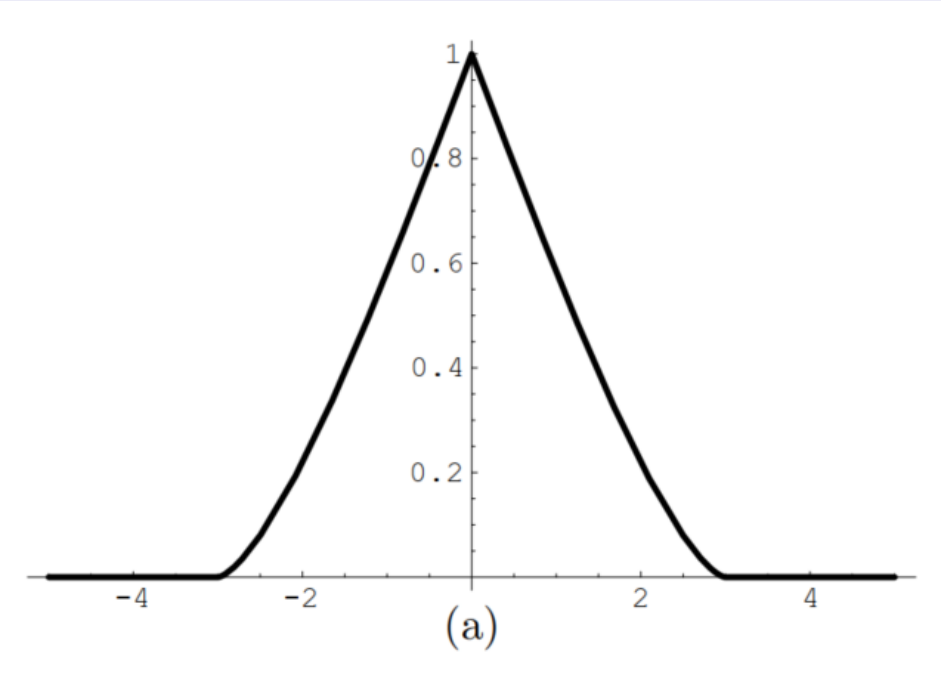
- The isotropic stationary kernels must fall off more quickly as the dimension d increases, as might be expected by examining the basis functions Ω_d. Those in R^∞ have the greatest restrictions placed on them.
- Isotropic stationary kernels that are positive definite in $\Re^d$ form a nested family of subspaces. When $d \to \infty$ the basis $\Omega_d(x)$ goes to $\exp(-x^2)$.
- *Schoenberg (1938)*: if β_d is the class of positive definite functions of the form given by *Bochner (1955)*, then the classes for all d have the property: $\beta_1 \subset \beta_2 \subset \ ... \ \subset \beta_d \subset ... \ \subset \beta_\infty$, so that as d is increased, the space of available functions is reduced.

Commonly used isotropic stationary kernels

$K_I(\|\boldsymbol{x}-\boldsymbol{z}\|)/K_I(0)$

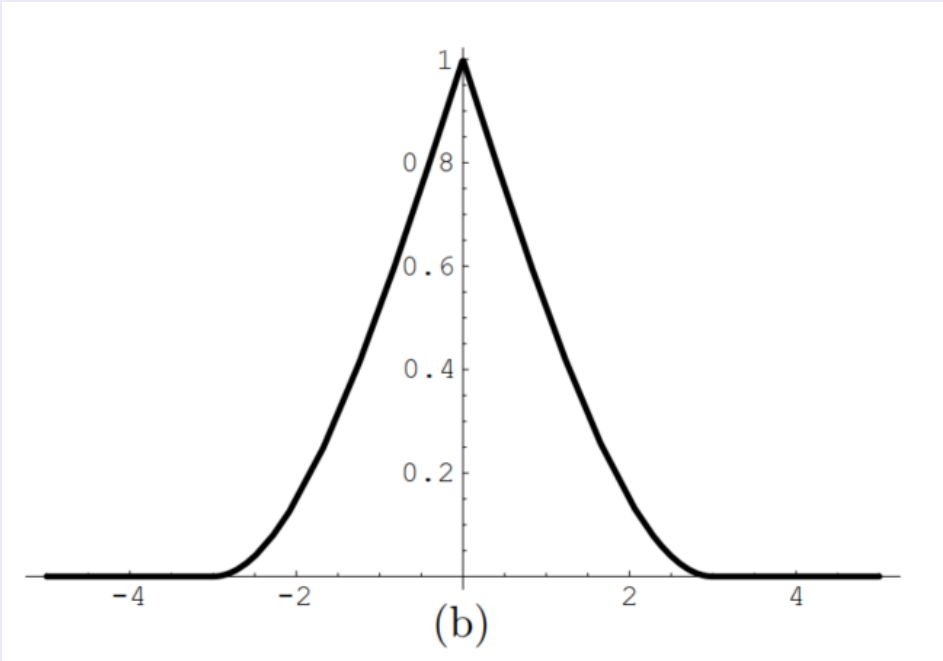
(a) Circular:

positive definite in $\Re^2$ $\frac{2}{\pi}\arccos\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)-\frac{2}{\pi}\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\sqrt{1-\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)^2}$ if $\|\boldsymbol{x}-\boldsymbol{z}\|<\theta$ zero otherwise



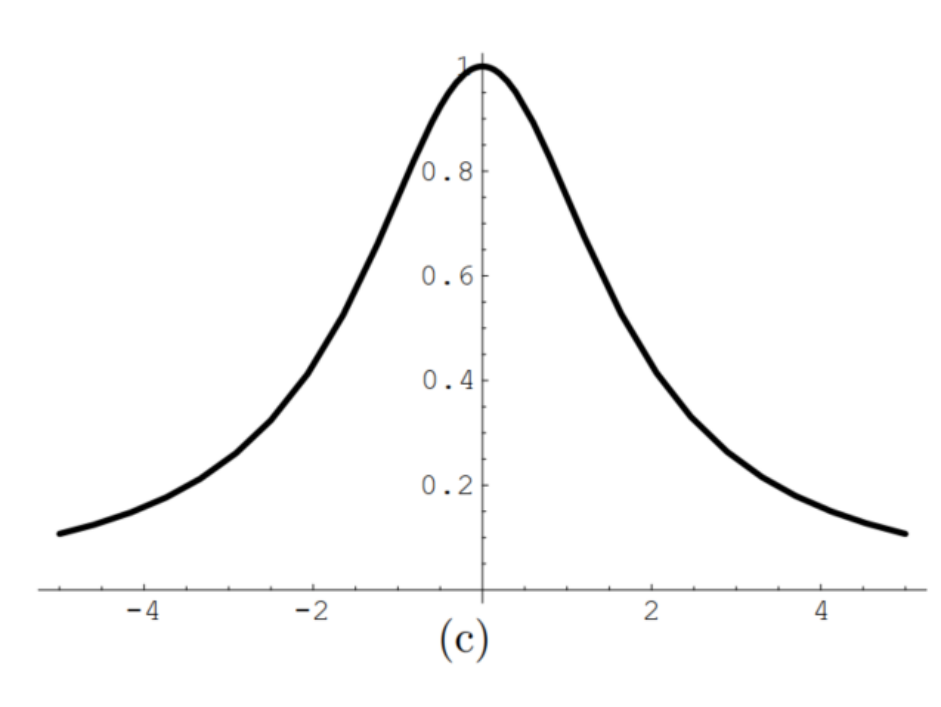
(b) Spherical:

positive definite in $\Re^3$ $1 - \frac{3}{2}\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta} + \frac{1}{2}\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)^3$ if $\|\boldsymbol{x}-\boldsymbol{z}\| < \theta$ zero otherwise



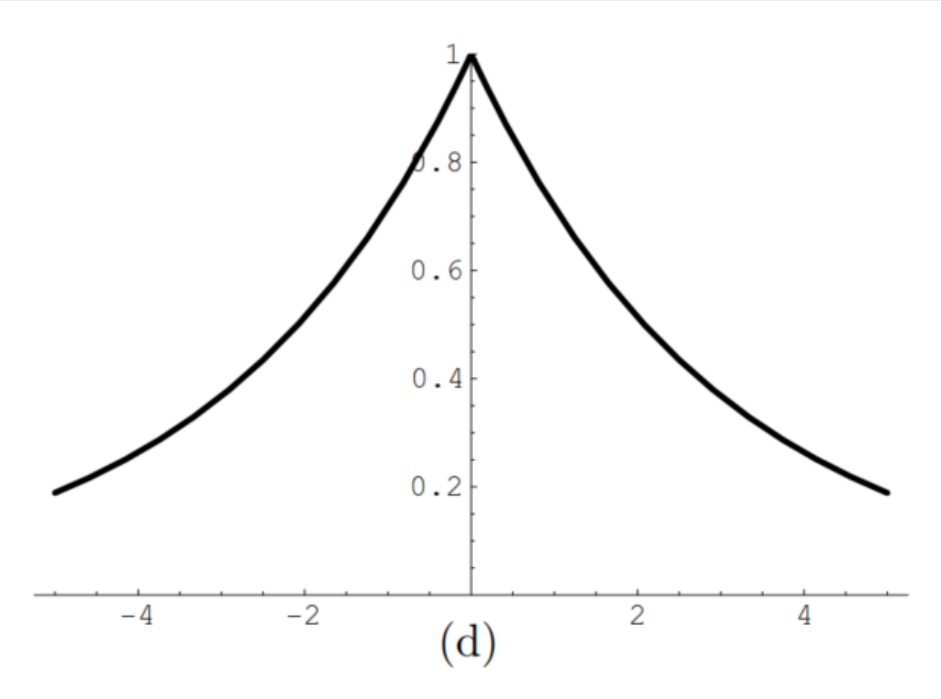
(c) Rational quadratic:

positive definite in $\Re^d$ $1 - \frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\|\boldsymbol{x}-\boldsymbol{z}\|^2+\theta}$



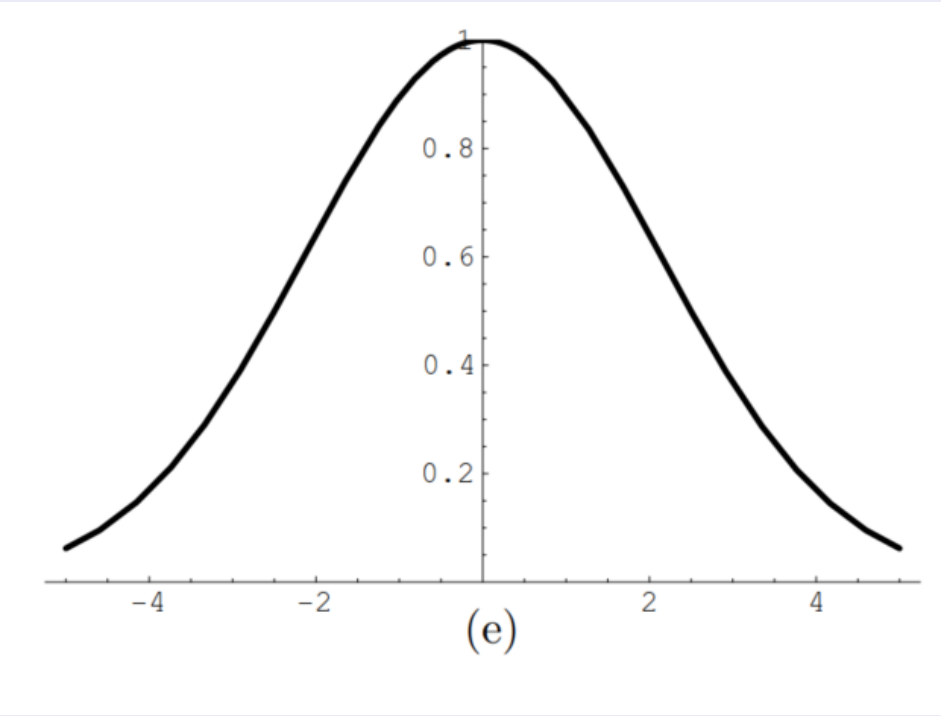
(d) Exponential:

positive definite in $\Re^d$ $\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)$



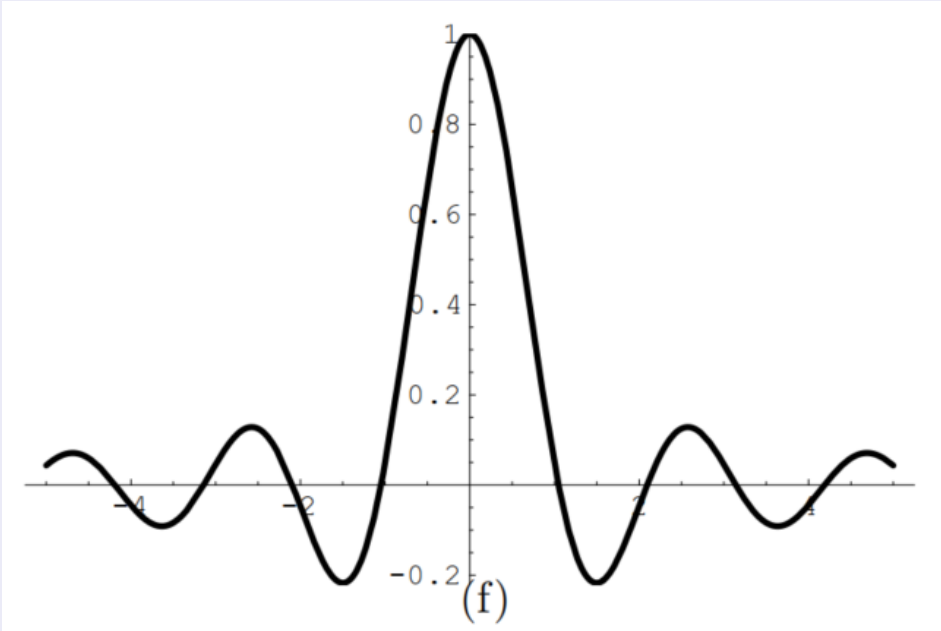
(e) Gaussian:

positive definite in $\Re^d$ $\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{\theta}\right)$



(f) Wave:

positive definite in $\Re^3$ $\frac{\theta}{\|\boldsymbol{x}-\boldsymbol{z}\|}\sin\left(\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)$



- $\theta > 0$ as parameter:
- The exponential kernel (d) is obtained from the spectral representation (11) with the spectral density: $f(\omega) = \frac{1}{\frac{\pi}{\theta} + \pi\theta\omega^2}$
- The Gaussian kernel (e) is obtained with the spectral density: $f(\omega) = \frac{\sqrt{\theta}}{2\sqrt{\pi}} \exp\left(- \frac{\theta\omega^2}{4} \right)$.
- Note also that the circular and spherical kernels have compact support. They have a linear behavior at the origin, which is also true for the exponential kernel.
- The rational quadratic, Gaussian, and wave kernels have a parabolic behavior at the origin. This indicates a different degree of smoothness.

Matern kernel

- The Matern kernel *(Matern, 1960)* has recently received considerable attention, because it allows to control the smoothness with a parameter ν.
- The Matern kernel is defined by:

$$K_I(\|\boldsymbol{x}-\boldsymbol{z}\|)/K_I(0) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\sqrt{\nu}\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right)^{\nu} H_\nu\left(\frac{2\sqrt{\nu}\|\boldsymbol{x}-\boldsymbol{z}\|}{\theta}\right) \quad (12)$$

where Γ is the Gamma function and H_ν is the modified Bessel function of the second kind of order ν.

- Note that the Matern kernel reduces to the exponential kernel for $\nu = 0.5$ and to the Gaussian kernel for $\nu \to \infty$.

Compactly supported kernels

- Compactly supported kernels are kernels that vanish whenever the distance between two examples $\boldsymbol{x}$ and $\boldsymbol{z}$ is larger than a certain cut-off distance, often called the range.
- For instance, the spherical kernel (b) is a compactly supported kernel since $K_I(\|\boldsymbol{x} - \boldsymbol{z}\|) = 0$ when $\|\boldsymbol{x} - \boldsymbol{z}\| \geq \theta$. This might prove a crucial advantage for certain applications dealing with massive data sets, because the corresponding Gram matrix G, $G_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, will be sparse.
- Then, linear systems involving the matrix G can be solved very efficiently using sparse linear algebra techniques, seen *Gilbert et al.(1992)*.

Example

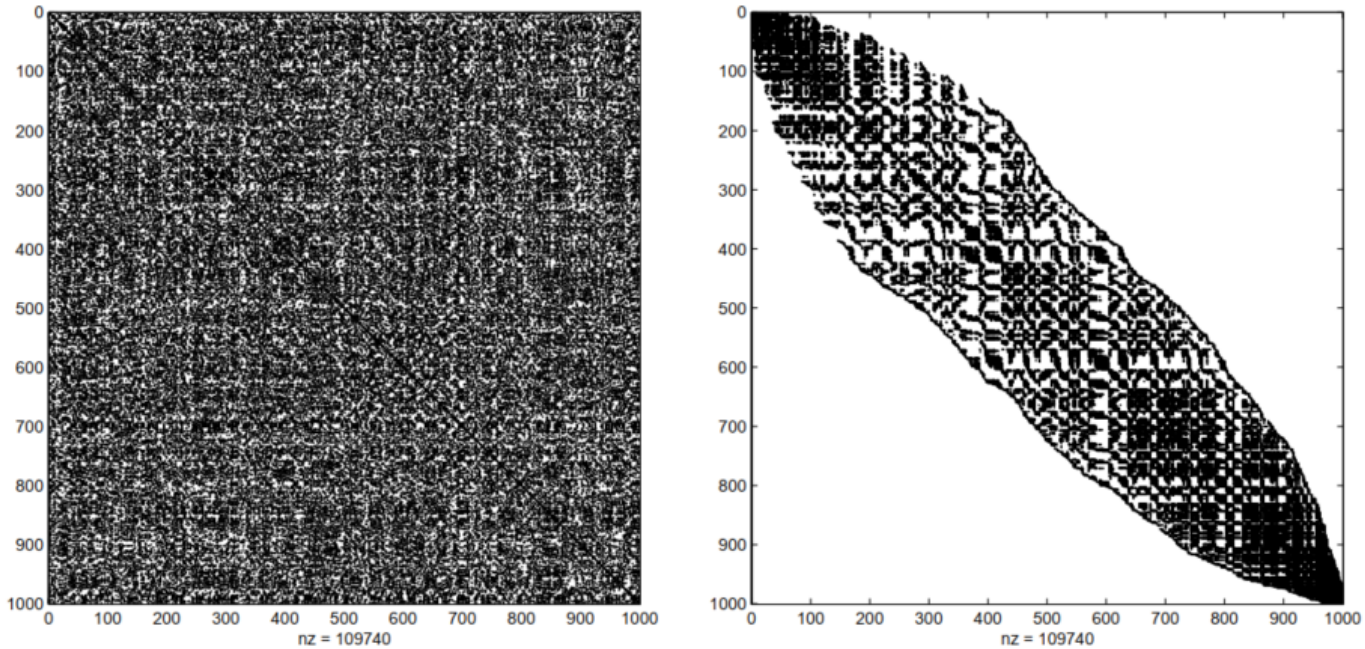


Figure 3: The Gram matrix for 1,000 examples uniformly distributed in the unit square, based on a spherical kernel with range $\theta = 0.2$: initial (left panel); after reordering (right panel).

Example (continued)

- The reordered Gram matrix has now a bandwidth of only 252 instead of 1,000 for the initial matrix, andimportant computational savings can be obtained.
- A compactly supported kernel of Matern type can be obtained by multiplying the kernel (12) by the kernel: $\max\left\{\left(1-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|}{\tilde{\theta}}\right)^{\tilde{\nu}},\ 0\right\}$, where $\tilde{\theta}>0$ and $\tilde{\nu}\geq(d+1)/2$, in order to insure positive definiteness.
- This product is a kernel by the property (3). Beware that it is not possible to simply "cut-off" a kernel in order to obtain a compactly supported one, because the result will not be positive definite in general.

Locally Stationary Kernels

(Silverman, 1957, 1959):

$$K(\boldsymbol{x},\ \boldsymbol{z}) = K_1\left(\frac{\boldsymbol{x}+\boldsymbol{z}}{2}\right) K_2(\boldsymbol{x}-\boldsymbol{z}) \qquad (13)$$

- where K_1 is a nonnegative function and K_2 is a stationary kernel. Note that if K_1 is a positive constant, then (13) reduces to a stationary kernel.
- Further impose that $K_2(\mathbf{0}) = 1$. The variable $(\boldsymbol{x}+\boldsymbol{z})/2$ has been chosen because of its suggestive meaning of the average or centroid of the examples $\boldsymbol{x}$ and $\boldsymbol{z}$.

- The variance is determined by:
  $K(\boldsymbol{x},\ \boldsymbol{x}) = K_1(\boldsymbol{x})K_2(\mathbf{0}) = K_1(\boldsymbol{x}) \quad (14)$ thus justifying the name of power schedule for $K_1(x)$, which describes the global structure.
- On the other hand, $K_2(\boldsymbol{x} - \boldsymbol{z})$ is invariant under shifts and thus describes the local structure. It can be obtained by considering:
  $K(\boldsymbol{x}/2,\ -\boldsymbol{x}/2) = K_1(\mathbf{0})K_2(\boldsymbol{x}) \quad (15)$

Properties

- Equations (14) and (15) imply that the kernel $K(\boldsymbol{x},\ \boldsymbol{z})$ defined by (13) is completely determinedby its values on the diagonal $\boldsymbol{x} = \boldsymbol{z}$ and antidiagonal $\boldsymbol{x} = -\boldsymbol{z}$, for:

$$K(\boldsymbol{x},\ \boldsymbol{z}) = \frac{K\left(\frac{(\boldsymbol{x}+\boldsymbol{z})}{2},\ \frac{(\boldsymbol{x}+\boldsymbol{z})}{2}\right) K\left(\frac{(\boldsymbol{x}-\boldsymbol{z})}{2},\ -\frac{(\boldsymbol{x}-\boldsymbol{z})}{2}\right)}{K(\mathbf{0},\ \mathbf{0})} \qquad (16)$$

- K_1 is invariant with respect to shifts parallel to the antidiagonal, whereas K_2 is invariant with respect to shifts parallel to the diagonal.
- These properties allow to find moment estimators of both K_1 and K_2 from a single realization of data, although the kernel is not stationary.

Exponentially convex kernels

- Another special class of locally stationary kernels is defined by kernels of the form: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_1(x+z) \quad (17)$ the so-called exponentially convex kernels *(Loeve, 1946, 1948)*. need to read more
- From (16), we see immediately that $K_1(\boldsymbol{x}+\boldsymbol{z}) \geq 0$. Actually, as notedby Loeve, any two-sided Laplace transform of a nonnegative function is an exponentially convex kernel.

Build locally stationary kernels

- A large class of locally stationary kernels can therefore be constructed by multiplying an exponentially convex kernel by a stationary kernel, since the product of two kernels is a kernel by the property (3).
- However, the following example is a locally stationary kernel in $\Re^1$ which is not the product of two kernels:
$$\exp[-a(x^2+z^2)] = \exp\left[-2a\left(\frac{x+z}{2}\right)^2\right] \exp\left[-\frac{a(x+z)^2}{2}\right],\ a>0 \quad (18)$$
since the first factor in the right side is a positive function without being a kernel, and the second factor is a kernel.

- Finally, with the positive definite Delta kernel $\delta(\boldsymbol{x}-\boldsymbol{z})$, which is equal to 1 if $\boldsymbol{x}=\boldsymbol{z}$ and 0 otherwise, the product: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_1\left(\frac{\boldsymbol{x}+\boldsymbol{z}}{2}\right)\delta(\boldsymbol{x}-\boldsymbol{z})$, is a locally stationary kernel, often called a *locally stationary white noise*.
- The spectral representation of locally stationary kernels has remarkable properties. Indeed, it can be written as *(Silverman, 1957)*: $K(\boldsymbol{x},\ \boldsymbol{z}) = \int_{\Re^d}\int_{\Re^d}\cos(\boldsymbol{\omega}_1^\top\boldsymbol{x}-\boldsymbol{\omega}_2^\top\boldsymbol{z})f_1(\frac{\boldsymbol{\omega}_1+\boldsymbol{\omega}_2}{2})f_2(\boldsymbol{\omega}_1-\boldsymbol{\omega}_2)d\boldsymbol{\omega}_1 d\boldsymbol{\omega}_2,$
- i.e. the spectral density $f_1(\frac{\boldsymbol{\omega}_1+\boldsymbol{\omega}_2}{2})f_2(\boldsymbol{\omega}_1-\boldsymbol{\omega}_2)$ is also a locally stationary kernel, and:

$$K_1(\boldsymbol{u}) = \int_{\Re^d}\cos(\boldsymbol{\omega}^\top\boldsymbol{u})f_2(\boldsymbol{\omega})d\boldsymbol{\omega}$$

$$K_2(\boldsymbol{v}) = \int_{\Re^d}\cos(\boldsymbol{\omega}^\top\boldsymbol{v})f_1(\boldsymbol{\omega})d\boldsymbol{\omega}$$

- i.e. K_1, f_2 and K_2, f_1 are Fourier transform pairs. For instance, to the locally stationary kernel (18) corresponds the spectral density:

$$f_1(\frac{\omega_1+\omega_2}{2})f_2(\omega_1-\omega_2) = \frac{1}{4\pi a}\exp\Big[-\frac{1}{2a}(\frac{(\omega_1+\omega_2)}{2})^2\Big]\exp\Big[-\frac{1}{8a}\frac{(\omega_1-\omega_2)^2}{2}\Big]$$

- This is immediately seen to be locally stationary since, except for a positive factor, it is of the form (18), with a replaced by $1/(4a)$.
- In particular, we can obtain a very rich family of locally stationary kernels by multiplying a Matern kernel (12) by an exponentially convex kernel (17). The resulting product is still a kernel by the property (3).

Nonstationary kernels

- The most general class of kernels is the one of nonstationary kernels, $K(\boldsymbol{x},\ \boldsymbol{z})$.
- For example, the polynomial kernel of degree p: $K(\boldsymbol{x},\ \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z})^p$, is a nonstationary kernel. The spectral representation of nonstationary kernels is very general. A nonstationary kernel $K(\boldsymbol{x},\ \boldsymbol{z})$ is positive definite in $\Re^d$ if and only if it has the form *(Yaglom, 1987)*: where F is a positive bounded symmetric measure.

$$K(\boldsymbol{x},\ \boldsymbol{z}) = \int_{\Re^d} \int_{\Re^d} \cos(\boldsymbol{\omega}_1^\top \boldsymbol{x} - \boldsymbol{\omega}_2^\top \boldsymbol{z}) F(d\boldsymbol{\omega}_1, d\boldsymbol{\omega}_2) \qquad (19)$$

- When the function $F(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ is concentrated on the diagonal $\omega_1 = \omega_2$, then (19) reduces to the spectral representation (10) of stationary kernels. Here again, many nonstationary kernels can be constructed with (19).

nonstationary kernel and spectral density

- Of interest are nonstationary kernels obtained from (19) with $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$ but with a spectral density that is not integrable in a neighborhood around the origin. Such kernels are referred to as generalized kernels *(Matheron, 1973)*.
- For instance, the Brownian motion generalized kernel corresponds to a spectral density $f(\omega) = 1/\|\omega\|^2$ *(Mandelbrot and Van Ness, 1968)*.
- A particular family of nonstationary kernels is the one of separable nonstationary kernels: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_1(\boldsymbol{x})K_2(\boldsymbol{z})$, where K_1 and K_2 are stationary kernels evaluated at the examples $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively.

Separable nonstationary

- Separable nonstationary kernels possess the property that their Gram matrix G, with $G_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, can be written as a tensor product (also called Kronecker product, see *Graham, 1981*) of two vectors defined by K_1 and K_2 respectively.
- This is especially useful to reduce computational burden when dealing with massive data sets. For instance, consider a set of l examples $\boldsymbol{x}_1, \ ..., \ \boldsymbol{x}_l$. The memory requirements for the computation of the Gram matrix is reduced from l^2 to $2l$ since it suffices to evaluate the vectors $\boldsymbol{a} = \left(K_1(\boldsymbol{x}_1), \ ..., \ K_1(\boldsymbol{x}_l)\right)^\top$ and $\boldsymbol{b} = \left(K_2(\boldsymbol{x}_1), \ ..., \ K_2(\boldsymbol{x}_l)\right)^\top$.
- We then have $G = \boldsymbol{a}\boldsymbol{b}^\top$. Such a computational reduction can be of crucial importance for certain applications involving very large training sets.

Main idea

- To find a new feature space where stationarity (see *Sampson and Guttorp, 1992*) or local stationarity (see *Genton and Perrin, 2001*) can be achieved.
- A nonstationary kernel $K(\boldsymbol{x},\ \boldsymbol{z})$ is stationary reducible if there exist a bijective deformation $\boldsymbol{\Phi}$ such that:
  $K(\boldsymbol{x},\ \boldsymbol{z}) = K_S^*(\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{z}))$ $\quad (20)$ where K_S^* is a stationary kernel.
- For example in $\Re^2$, the nonstationary kernel defined by:
  $K(\boldsymbol{x},\ \boldsymbol{z}) = \frac{\|\boldsymbol{x}\|+\|\boldsymbol{z}\|-\|\boldsymbol{z}-\boldsymbol{x}\|}{2\sqrt{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}}$ $\quad (21)$ is stationary reducible with the deformation: $\boldsymbol{\Phi}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\ln\left(\sqrt{x_1^2 + x_2^2}\right),\ \arctan(x_2/x_1) \right)^\top$,
  yielding the stationary kernel:

$$K_S^*(\boldsymbol{u}_1, \boldsymbol{u}_2) = \cosh(\frac{\boldsymbol{u}_1}{2}) - \sqrt{\frac{\cosh(\frac{\boldsymbol{u}1}{2}) - \cos(\boldsymbol{u}_2)}{2}} \qquad (22)$$

Differentiable condition

- Effectively, it is straightforward to check with some algebra that (22) evaluated at:
  $\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{z}) = \left(\ln\left(\frac{\|\boldsymbol{x}\|}{\|\boldsymbol{z}\|}\right),\ \arctan\left(\frac{x_2}{x_1}\right) - \arctan\left(\frac{z_2}{z_1}\right) \right)^\top$ yields the kernel (21).
- Specifically, if $\boldsymbol{\Phi}$ and its inverse are differentiable in $\Re^d$, and $K(\boldsymbol{x},\ \boldsymbol{z})$ is continuously differentiable for $\boldsymbol{x} \neq \boldsymbol{y}$, then K satisfies (20) if and only if:

$$D_{\boldsymbol{x}} K(\boldsymbol{x},\ \boldsymbol{z}) Q_{\boldsymbol{\Phi}}^{-1}(\boldsymbol{x}) + D_{\boldsymbol{z}} K(\boldsymbol{x},\ \boldsymbol{z}) Q_{\boldsymbol{\Phi}}^{-1}(\boldsymbol{z}) = \mathbf{0},\ \boldsymbol{x} \neq \boldsymbol{y} \quad (23)$$

where $Q_{\boldsymbol{\Phi}}$ is the Jacobian of $\boldsymbol{\Phi}$ and $D_{\boldsymbol{x}}$ denotes the partial derivatives operator with respect to $\boldsymbol{x}$.

Not all nonstationary are reducible

- Unfortunately, not all nonstationary kernels can be reduced to stationarity through a deformation $\boldsymbol{\Phi}$. Consider for instance the kernel in $\Re^1$: $K(\boldsymbol{x},\ \boldsymbol{z}) = \exp(2 - x^6 - z^6) \quad (24)$, which is positive definite as can be seen from (6).
- It is obvious that $K(\boldsymbol{x},\ \boldsymbol{z})$ does not satisfy Equation (23) and thus is not stationary reducible. This is the motivation of *Genton and Perrin (2001)* to extend the model (20) to locally stationary kernels.
- We say that a nonstationary kernel K is locally stationary reducible if there exists a bijective deformation $\boldsymbol{\Phi}$ such that: $K(\boldsymbol{x},\ \boldsymbol{z}) = K_1\left(\frac{\boldsymbol{\Phi}(\boldsymbol{x})+\boldsymbol{\Phi}(\boldsymbol{z})}{2}\right) K_2(\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{z})) \quad (25)$, where K_1 is a nonnegative function and K_2 is a stationary kernel.

- Note that if K_1 is a positive constant, then Equation (25) reduces to the model (20). *Genton and Perrin (2001)* characterize such transformations $\mathbf{\Phi}$.
- For instance, the nonstationary kernel (24) can be reduced to a locally stationary kernel with the transformation: $\mathbf{\Phi}(x) = \frac{x^3}{3} - \frac{1}{3}$ (26), yielding: $K_1(u) = \exp(-18u^2 - 12u)$ (27) and $K_2(v) = \exp\left(-\frac{9}{2}v^2\right)$ (28)

- Here again, it can easily be checked from (27), (28), and (26) that: $K_1\left(\frac{\mathbf{\Phi}(x)+\mathbf{\Phi}(z)}{2}\right)K_2(\mathbf{\Phi}(x)-\mathbf{\Phi}(z)) = \exp(2-x^6-z^6)$.
- Of course, it is possible to construct nonstationary kernels that are neither stationary reducible nor locally stationary reducible.
- Actually, the familiar class of polynomial kernels of degree p, $K(\boldsymbol{x},\ \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z})^p$, cannot be reduced to stationarity or local stationarity with a bijective transformation $\mathbf{\Phi}$.

Summary

- Kernels introduced in this paper
  - stationary (anisotropic/isotropic/compactly supported)
  - locally stationary
  - nonstationary
  - separable nonstationary kernels
- Each class has its own particular properties and spectral representation (allows for the design of many new kernels in each class).
  - Note that kernels from the classes presented in this paper can be combined indefinitely by using the properties (2)-(9).
  - This should prove useful to researchers designing new kernels and algorithms for machine learning.
  - In particular, the reducibility of nonstationary kernels to simpler kernels which are stationary or locally stationary suggests interesting applications.

- Nonstationary kernel
- Generalized kernel
- Separable kernel
- Reducible kernel
- Locally Stationary kernel
- Stationary
  - isotropic
  - compactly supported
- **Stationary kernel** is a friend just like **Exponential family distribution**
- The goal is to find the "Minimal Sufficient Statistics", although we cannot always find one.