Journal of Machine Learning Research 2 (2001) 299-312 Submitted 3/01; Published 12/01

Classes of Kernels for Machine Learning: A Statistics Perspective

Marc G. Genton GENTON@STAT.NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203, USA

Editors: Nello Cristianini, John Shawe-Taylor, Robert Williamson

Abstract

In this paper, we present classes of kernels for machine learning from a statistics perspective. Indeed, kernels are positive definite functions and thus also covariances. After discussing key properties of kernels, as well as a new formula to construct kernels, we present several important classes of kernels: anisotropic stationary kernels, isotropic stationary kernels, compactly supported kernels, locally stationary kernels, nonstationary kernels, and separable nonstationary kernels. Compactly supported kernels and separable nonstationary kernels are of prime interest because they provide a computational reduction for kernel-based methods. We describe the spectral representation of the various classes of kernels and conclude with a discussion on the characterization of nonlinear maps that reduce nonstationary kernels to either stationarity or local stationarity.

Keywords: Anisotropic, Compactly Supported, Covariance, Isotropic, Locally Stationary, Nonstationary, Reducible, Separable, Stationary

1. Introduction

Recently, the use of kernels in learning systems has received considerable attention. The main reason is that kernels allow to map the data into a high dimensional feature space in order to increase the computational power of linear machines (see for example Vapnik, 1995, 1998, Cristianini and Shawe-Taylor, 2000). Thus, it is a way of extending linear hypotheses to nonlinear ones, and this step can be performed implicitly. Support vector machines, kernel principal component analysis, kernel Gram-Schmidt, Bayes point machines, Gaussian processes, are just some of the algorithms that make crucial use of kernels for problems of classification, regression, density estimation, and clustering. In this paper, we present classes of kernels for machine learning from a statistics perspective. We discuss simple methods to design kernels in each of those classes and describe the algebra associated with kernels.

The kinds of kernel K we will be interested in are such that for all examples $\mathbf{x}$ and $\mathbf{z}$ in an input space $X \subset \mathbb{R}^d$:

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

where ϕ is a nonlinear (or sometimes linear) map from the input space X to the feature space $\mathcal{F}$, and $\langle \cdot, \cdot \rangle$ is an inner product. Note that kernels can be defined on more general input spaces X, see for instance Aronszajn (1950). In practice, the kernel K is usually defined directly, thus implicitly defining the map ϕ and the feature space $\mathcal{F}$. It is therefore important to be able to design new kernels. Clearly, from the symmetry of the inner product, a kernel must be symmetric:

$$K(\mathbf{x},\mathbf{z}) = K(\mathbf{z},\mathbf{x}),$$

and also satisfy the Cauchy-Schwartz inequality:

$$K^2(\mathbf{x},\mathbf{z}) \leq K(\mathbf{x},\mathbf{x})K(\mathbf{z},\mathbf{z}).$$

However, this is not sufficient to guarantee the existence of a feature space. Mercer (1909) showed that a necessary and sufficient condition for a symmetric function $K(\mathbf{x},\mathbf{z})$ to be a kernel is that it be positive definite. This means that for any set of examples $\mathbf{x}_1,\ldots,\mathbf{x}_l$ and any set of real numbers $\lambda_1,\ldots,\lambda_l$, the function K must satisfy:

$$\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j K(\mathbf{x}_i,\mathbf{x}_j) \geq 0. \tag{1}$$

Symmetric positive definite functions are called covariances in the statistics literature. Hence kernels are essentially covariances, and we propose a statistics perspective on the design of kernels. It is simple to create new kernels from existing kernels because positive definite functions have a pleasant algebra, and we list some of their main properties below. First, if K_1, K_2 are two kernels, and a_1, a_2 are two positive real numbers, then:

$$K(\mathbf{x},\mathbf{z}) = a_1K_1(\mathbf{x},\mathbf{z}) + a_2K_2(\mathbf{x},\mathbf{z}), \tag{2}$$

is a kernel. This result implies that the family of kernels is a convex cone. The multiplication of two kernels K_1 and K_2 yields a kernel:

$$K(\mathbf{x},\mathbf{z}) = K_1(\mathbf{x},\mathbf{z})K_2(\mathbf{x},\mathbf{z}). \tag{3}$$

Properties (2) and (3) imply that any polynomial with positive coefficients, $pol^+(x) = \{\sum_{i=1}^{n}\alpha_i x^i | n \in \mathbb{N},\ \alpha_1,\ldots,\alpha_n \in \mathbb{R}^+\}$, evaluated at a kernel K_1, yields a kernel:

$$K(\mathbf{x},\mathbf{z}) = pol^+(K_1(\mathbf{x},\mathbf{z})). \tag{4}$$

In particular, we have that:

$$K(\mathbf{x},\mathbf{z}) = \exp(K_1(\mathbf{x},\mathbf{z})), \tag{5}$$

is a kernel by taking the limit of the series expansion of the exponential function. Next, if g is a real-valued function on X, then

$$K(\mathbf{x},\mathbf{z}) = g(\mathbf{x})g(\mathbf{z}), \tag{6}$$

is a kernel. If ψ is an $\mathbb{R}^p$-valued function on X and K_3 is a kernel on $\mathbb{R}^p \times \mathbb{R}^p$, then:

$$K(\mathbf{x},\mathbf{z}) = K_3(\psi(\mathbf{x}),\psi(\mathbf{z})), \tag{7}$$

is also a kernel. Finally, if A is a positive definite matrix of size $d \times d$, then:

$$K(\mathbf{x},\mathbf{z}) = \mathbf{x}^T A\mathbf{z}, \tag{8}$$

is a kernel. The results (2)-(8) can easily be derived from (1), see also Cristianini and Shawe-Taylor (2000). The following property can be used to construct kernels and seems not to be known in the machine learning literature. Let h be a real-valued function on X, positive, with minimum at 0 (that is, h is a variance function). Then:

$$K(\mathbf{x}, \mathbf{z}) = \frac{1}{4}\Big[h(\mathbf{x}+\mathbf{z}) - h(\mathbf{x}-\mathbf{z})\Big], \tag{9}$$

is a kernel. The justification of (9) comes from the following identity for two random variables Y_1 and Y_2: Covariance(Y_1,Y_2)=[Variance(Y_1 + Y_2)$-$Variance(Y_1 $-$ Y_2)]/4. For instance, consider the function $h(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. From (9), we obtain the kernel:

$$K(\mathbf{x}, \mathbf{z}) = \frac{1}{4}\Big[(\mathbf{x}+\mathbf{z})^T(\mathbf{x}+\mathbf{z}) - (\mathbf{x}-\mathbf{z})^T(\mathbf{x}-\mathbf{z})\Big] = \mathbf{x}^T\mathbf{z}.$$

The remainder of the paper is set up as follows. In Section 2, 3, and 4, we discuss respectively the class of stationary, locally stationary, and nonstationary kernels. Of particular interest are the classes of compactly supported kernels and separable nonstationary kernels because they reduce the computational burden of kernel-based methods. For each class of kernels, we present their spectral representation and show how it can be used to design many new kernels. Section 5 addresses the reducibility of nonstationary kernels to stationarity or local stationarity, and we conclude the paper in Section 6.

2. Stationary Kernels

A stationary kernel is one which is translation invariant:

$$K(\mathbf{x}, \mathbf{z}) = K_S(\mathbf{x}-\mathbf{z}),$$

that is, it depends only on the lag vector separating the two examples $\mathbf{x}$ and $\mathbf{z}$, but not on the examples themselves. Such a kernel is sometimes referred to as anisotropic stationary kernel, in order to emphasize the dependence on both the direction and the length of the lag vector. The assumption of stationarity has been extensively used in time series (see for example Brockwell and Davis, 1991) and spatial statistics (see for example Cressie, 1993) because it allows for inference on K based on all pairs of examples separated by the same lag vector. Many stationary kernels can be constructed from their spectral representation derived by Bochner (1955). He proved that a stationary kernel $K_S(\mathbf{x}-\mathbf{z})$ is positive definite in $\mathbb{R}^d$ if and only if it has the form:

$$K_S(\mathbf{x}-\mathbf{z}) = \int_{\mathbb{R}^d} \cos\big(\boldsymbol{\omega}^T(\mathbf{x}-\mathbf{z})\big) F(d\boldsymbol{\omega}), \tag{10}$$

where F is a positive finite measure. The quantity $F/K_S(\mathbf{0})$ is called the spectral distribution function. Note that (10) is simply the Fourier transform of F. Cressie and Huang (1999) and Gneiting (2002b) use (10) to derive nonseparable space-time stationary kernels, see also Christakos (2000) for illustrative examples.

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:

$$K(\mathbf{x},\mathbf{z}) = K_I(\|\mathbf{x}-\mathbf{z}\|).$$

The spectral representation of isotropic stationary kernels has been derived from Bochner's theorem (Bochner, 1955) by Yaglom (1957):

$$K_I(\|\mathbf{x}-\mathbf{z}\|) = \int_0^\infty \Omega_d\big(\omega\|\mathbf{x}-\mathbf{z}\|\big)F(d\omega), \qquad (11)$$

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 $\mathbb{R}^d$. 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. Some familiar examples of Ω_d are $\Omega_1(x) = \cos(x)$, $\Omega_2(x) = J_0(x)$, and $\Omega_3(x) = \sin(x)/x$. Here again, by choosing a nondecreasing bounded function F (or its derivative f), we can derive the corresponding kernel from (11). For instance in $\mathbb{R}^1$, with the spectral density $f(\omega) = (1-\cos(\omega))/(\pi\omega^2)$, we derive the triangular kernel:

$$\begin{aligned} K_I(x-z) &= \int_0^\infty \cos(\omega|x-z|)\frac{1-\cos(\omega)}{\pi\omega^2}d\omega \\ &= \frac{1}{2}(1-|x-z|)^+, \end{aligned}$$

where $(x)^+ = \max(x,0)$ (see Figure 1). Note that an isotropic stationary kernel obtained with Ω_d is positive definite in $\mathbb{R}^d$ and in lower dimensions, but not necessarily in higher dimensions. For example, the kernel $K_I(x-z) = (1-|x-z|)^+/2$ is positive definite in $\mathbb{R}^1$ but not in $\mathbb{R}^2$, see Cressie (1993, p.84) for a counterexample. It is interesting to remark from (11) that an isotropic stationary kernel has a lower bound (Stein, 1999):

$$K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0) \geq \inf_{x\geq 0} \Omega_d(x),$$

thus yielding:

$$\begin{aligned} K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0) &\geq -1 && \text{in } \mathbb{R}^1 \\ K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0) &\geq -0.403 && \text{in } \mathbb{R}^2 \\ K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0) &\geq -0.218 && \text{in } \mathbb{R}^3 \\ K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0) &\geq 0 && \text{in } \mathbb{R}^\infty. \end{aligned}$$

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 $\mathbb{R}^\infty$ have the greatest restrictions placed on them. Isotropic stationary kernels that are positive definite in $\mathbb{R}^d$ form a nested family of subspaces. When $d \to \infty$ the basis $\Omega_d(x)$ goes to $\exp(-x^2)$. Schoenberg

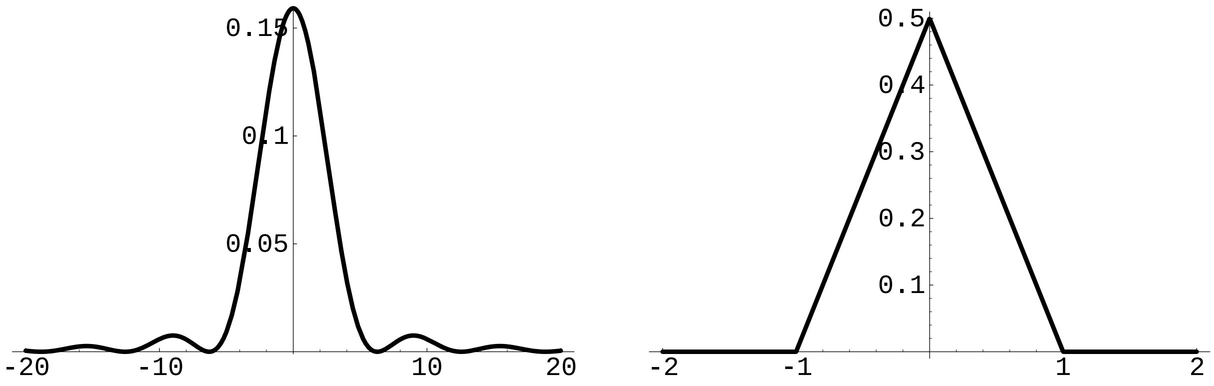

Figure 1: The spectral density $f(\omega) = (1 - \cos(\omega))/(\pi\omega^2)$ (left) and its corresponding isotropic stationary kernel $K_I(x - z) = (1 - |x - z|)^+/2$ (right).

(1938) proved that 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:

$$\beta_1 \supset \beta_2 \supset \cdots \supset \beta_d \supset \cdots \supset \beta_\infty,$$

so that as d is increased, the space of available functions is reduced. Only functions with the basis $\exp(-x^2)$ are contained in all the classes. The positive definite requirement imposes a smoothness condition on the basis as the dimension d is increased. Several criteria to check the positive definiteness of stationary kernels can be found in Christakos (1984). Further isotropic stationary kernels defined with non-Euclidean norms have recently been discussed by Christakos and Papanicolaou (2000).

From the spectral representation (11), we can construct many isotropic stationary kernels. Some of the most commonly used are depicted in Figure 2. They are defined by the equations listed in Table 1, where $\theta > 0$ is a parameter. As an illustration, the exponential kernel (d) is obtained from the spectral representation (11) with the spectral density:

$$f(\omega) = \frac{1}{\frac{\pi}{\theta} + \pi\theta\omega^2},$$

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

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

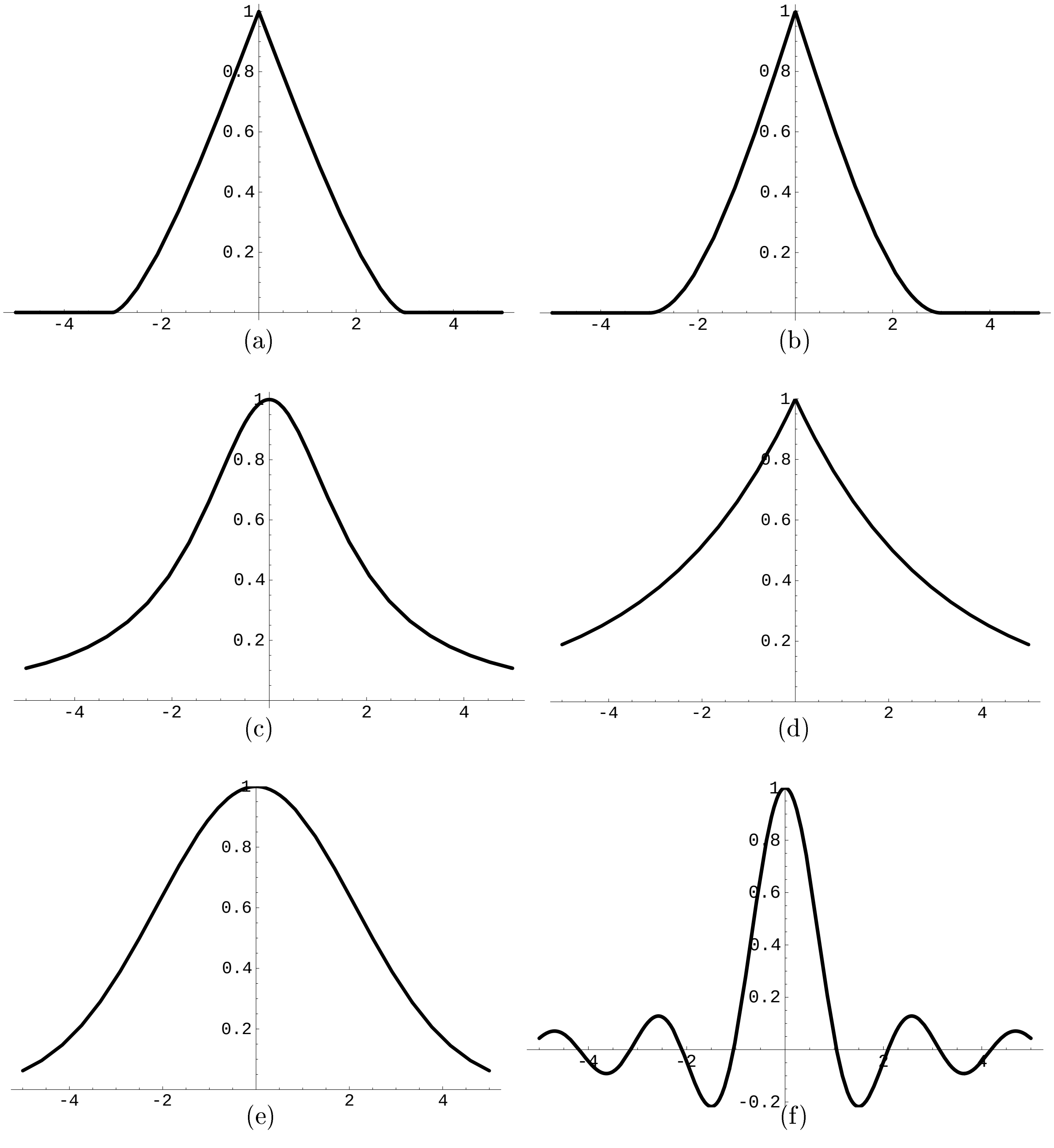

Figure 2: Some isotropic stationary kernels: (a) circular; (b) spherical; (c) rational quadratic; (d) exponential; (e) Gaussian; (f) wave.

where Γ is the Gamma function and H_ν is the modified Bessel function of the second kind of order ν. Note that the Matérn kernel reduces to the exponential kernel for $\nu = 0.5$ and

Name of kernel	$K_I(\|\mathbf{x}-\mathbf{z}\|)/K_I(0)$
(a) **Circular** positive definite in $\mathbb{R}^2$	$\frac{2}{\pi}\arccos\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\right) - \frac{2}{\pi}\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\sqrt{1-\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\right)^2}$ if $\|\mathbf{x}-\mathbf{z}\| < \theta$ zero otherwise
(b) **Spherical** positive definite in $\mathbb{R}^3$	$1-\frac{3}{2}\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}+\frac{1}{2}\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\right)^3$ if $\|\mathbf{x}-\mathbf{z}\| < \theta$ zero otherwise
(c) **Rational quadratic** positive definite in $\mathbb{R}^d$	$1-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\|\mathbf{x}-\mathbf{z}\|^2+\theta}$
(d) **Exponential** positive definite in $\mathbb{R}^d$	$\exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\right)$
(e) **Gaussian** positive definite in $\mathbb{R}^d$	$\exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\theta}\right)$
(f) **Wave** positive definite in $\mathbb{R}^3$	$\frac{\theta}{\|\mathbf{x}-\mathbf{z}\|}\sin\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\theta}\right)$

Table 1: Some commonly used isotropic stationary kernels.

to the Gaussian kernel for $\nu \to \infty$. Therefore, the Matérn kernel includes a large class of kernels and will prove very useful for applications because of this flexibility.

Compactly supported kernels are kernels that vanish whenever the distance between two examples $\mathbf{x}$ and $\mathbf{z}$ is larger than a certain cut-off distance, often called the range. For instance, the spherical kernel (b) is a compactly supported kernel since $K_I(\|\mathbf{x}-\mathbf{z}\|) = 0$ when $\|\mathbf{x}-\mathbf{z}\| \geq \theta$. This might prove a crucial advantage for certain applications dealing with massive data sets, because the corresponding Gram matrix G, whose ij-th element is $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, will be sparse. Then, linear systems involving the matrix G can be solved very efficiently using sparse linear algebra techniques, see for example Gilbert et al. (1992). As an illustrative example in $\mathbb{R}^2$, consider 1,000 examples, uniformly distributed in the unit square. Suppose that a spherical kernel (b) is used with a range of $\theta = 0.2$. The corresponding Gram matrix contains 1,000,000 entries, of which only 109,740 are not equal to zero, and is represented in the left panel of Figure 3 (black dots represent nonzero entries). The entries of the Gram matrix can be reordered, for instance with a sparse reverse Cuthill-McKee algorithm (see Gilbert et al., 1992), in order to have the nonzero elements closer to the diagonal. The result is displayed in the right panel of Figure 3. The reordered Gram matrix has now a bandwidth of only 252 instead of 1,000 for the initial matrix, and important computational savings can be obtained. Of course, if the

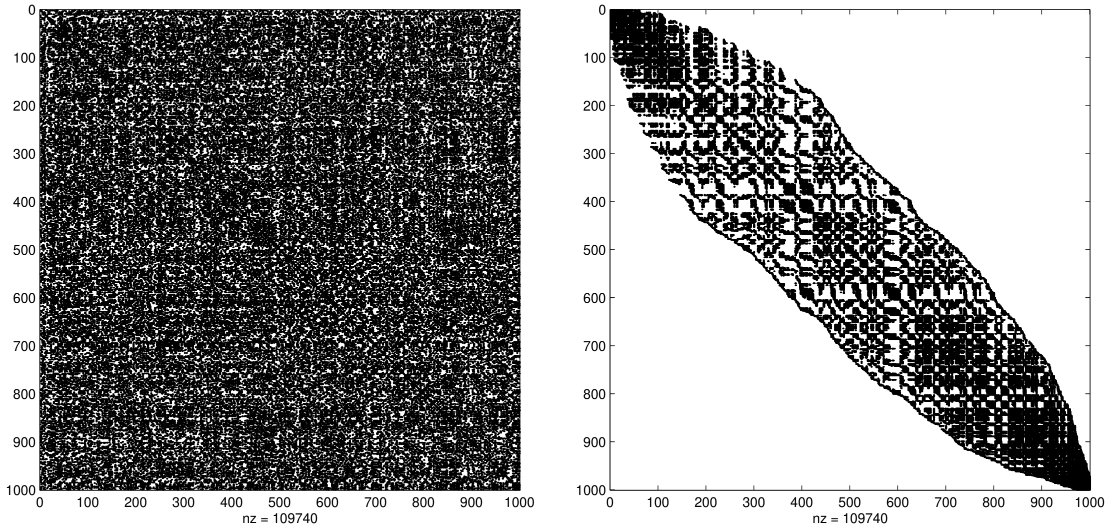

Figure 3: The Gram matrix for 1,000 examples uniformly distributed in the unit square, based on a spherical kernel with range $\theta = 0.2$: initial (left panel); after reordering (right panel).

spherical and the circular kernels would be the only compactly supported kernels available, this technique would be limited. Fortunately, large classes of compactly supported kernels can be constructed, see for example Gneiting (2002a) and references therein. A compactly supported kernel of Matérn type can be obtained by multiplying the kernel (12) by the kernel:

$$\max\left\{\left(1 - \frac{\|\mathbf{x}-\mathbf{z}\|}{\tilde{\theta}}\right)^{\tilde{\nu}}, 0\right\},$$

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.

3. Locally Stationary Kernels

A simple departure from the stationary kernels discussed in the previous section is provided by locally stationary kernels (Silverman, 1957, 1959):

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

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. Thus, the class of locally stationary kernels has the desirable property of including stationary kernels as a special case. Because the product of K_1 and K_2 is defined only up to a multiplicative positive constant, we further impose that $K_2(\mathbf{0}) = 1$. The variable $(\mathbf{x}+\mathbf{z})/2$ has been chosen because of its suggestive meaning of the average or centroid of the examples $\mathbf{x}$ and $\mathbf{z}$. The variance is determined by:

$$K(\mathbf{x},\mathbf{x}) = K_1(\mathbf{x})K_2(\mathbf{0}) = K_1(\mathbf{x}), \tag{14}$$

thus justifying the name of power schedule for $K_1(\mathbf{x})$, which describes the global structure. On the other hand, $K_2(\mathbf{x}-\mathbf{z})$ is invariant under shifts and thus describes the local structure. It can be obtained by considering:

$$K(\mathbf{x}/2, -\mathbf{x}/2) = K_1(\mathbf{0})K_2(\mathbf{x}). \tag{15}$$

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

$$K(\mathbf{x},\mathbf{z}) = \frac{K((\mathbf{x}+\mathbf{z})/2,(\mathbf{x}+\mathbf{z})/2)K((\mathbf{x}-\mathbf{z})/2,-(\mathbf{x}-\mathbf{z})/2)}{K(\mathbf{0},\mathbf{0})}. \tag{16}$$

Thus, we see that 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.

We already mentioned that stationary kernels are locally stationary. Another special class of locally stationary kernels is defined by kernels of the form:

$$K(\mathbf{x},\mathbf{z}) = K_1(\mathbf{x}+\mathbf{z}), \tag{17}$$

the so-called exponentially convex kernels (Loève, 1946, 1948). From (16), we see immediately that $K_1(\mathbf{x}+\mathbf{z}) \geq 0$. Actually, as noted by Loève, any two-sided Laplace transform of a nonnegative function is an exponentially convex kernel. 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 $\mathbb{R}^1$ which is not the product of two kernels:

$$\exp\Big[-a(x^2+z^2)\Big] = \exp\Big[-2a((x+z)/2)^2\Big]\,\exp\Big[-a(x-z)^2/2\Big],\ a>0, \tag{18}$$

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 $\delta(\mathbf{x}-\mathbf{z})$, which is equal to 1 if $\mathbf{x}=\mathbf{z}$ and 0 otherwise, the product:

$$K(\mathbf{x},\mathbf{z}) = K_1\Big(\frac{\mathbf{x}+\mathbf{z}}{2}\Big)\delta(\mathbf{x}-\mathbf{z}),$$

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):

$$K(\mathbf{x},\mathbf{z}) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \cos\big(\boldsymbol{\omega}_1^T\mathbf{x}-\boldsymbol{\omega}_2^T\mathbf{z}\big) f_1\Big(\frac{\boldsymbol{\omega}_1+\boldsymbol{\omega}_2}{2}\Big) f_2(\boldsymbol{\omega}_1-\boldsymbol{\omega}_2)d\boldsymbol{\omega}_1 d\boldsymbol{\omega}_2,$$

i.e. the spectral density $f_1\left(\frac{\boldsymbol{\omega}_1+\boldsymbol{\omega}_2}{2}\right)f_2(\boldsymbol{\omega}_1-\boldsymbol{\omega}_2)$ is also a locally stationary kernel, and:

$$
\begin{aligned}
K_1(\mathbf{u}) &= \int_{\mathbb{R}^d} \cos(\boldsymbol{\omega}^T\mathbf{u}) f_2(\boldsymbol{\omega}) d\boldsymbol{\omega}, \\
K_2(\mathbf{v}) &= \int_{\mathbb{R}^d} \cos(\boldsymbol{\omega}^T\mathbf{v}) f_1(\boldsymbol{\omega}) d\boldsymbol{\omega},
\end{aligned}
$$

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:

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

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

4. Nonstationary Kernels

The most general class of kernels is the one of nonstationary kernels, which depend explicitly on the two examples $\mathbf{x}$ and $\mathbf{z}$:

$$K(\mathbf{x}, \mathbf{z}).$$

For example, the polynomial kernel of degree p:

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^p,$$

is a nonstationary kernel. The spectral representation of nonstationary kernels is very general. A nonstationary kernel $K(\mathbf{x}, \mathbf{z})$ is positive definite in $\mathbb{R}^d$ if and only if it has the form (Yaglom, 1987):

$$K(\mathbf{x}, \mathbf{z}) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \cos\left(\boldsymbol{\omega}_1^T\mathbf{x} - \boldsymbol{\omega}_2^T\mathbf{z}\right) F(d\boldsymbol{\omega}_1, d\boldsymbol{\omega}_2), \tag{19}$$

where F is a positive bounded symmetric measure. When the function $F(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ is concentrated on the diagonal $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$, then (19) reduces to the spectral representation (10) of stationary kernels. Here again, many nonstationary kernels can be constructed with (19). 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 $f(\boldsymbol{\omega}) = 1/\|\boldsymbol{\omega}\|^2$ (Mandelbrot and Van Ness, 1968).

A particular family of nonstationary kernels is the one of separable nonstationary kernels:

$$K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x})K_2(\mathbf{z}),$$

where K_1 and K_2 are stationary kernels evaluated at the examples $\mathbf{x}$ and $\mathbf{z}$ respectively. The resulting product is a kernel by the property (3) in Section 1. Separable nonstationary kernels possess the property that their Gram matrix G, whose ij-th element is $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, 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 $\mathbf{x}_1, \ldots, \mathbf{x}_l$. The memory requirements fot the computation of the Gram matrix is then reduced from l^2 to $2l$ since it suffices to evaluate the vectors $\mathbf{a} = (K_1(\mathbf{x}_1), \ldots, K_1(\mathbf{x}_l))^T$ and $\mathbf{b} = (K_2(\mathbf{x}_1), \ldots, K_2(\mathbf{x}_l))^T$. We then have $G = \mathbf{a}\mathbf{b}^T$. Such a computational reduction can be of crucial importance for certain applications involving very large training sets.

5. Reducible Kernels

In this section, we discuss the characterization of nonlinear maps that reduce nonstationary kernels to either stationarity or local stationarity. The main idea is to find a new feature space where stationarity (see Sampson and Guttorp, 1992) or local stationarity (see Genton and Perrin, 2001) can be achieved. We say that a nonstationary kernel $K(\mathbf{x}, \mathbf{z})$ is stationary reducible if there exist a bijective deformation $\mathbf{\Phi}$ such that:

$$K(\mathbf{x}, \mathbf{z}) = K_S^*(\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{z})), \tag{20}$$

where K_S^* is a stationary kernel. For example in $\mathbb{R}^2$, the nonstationary kernel defined by:

$$K(\mathbf{x}, \mathbf{z}) = \frac{\|\mathbf{x}\| + \|\mathbf{z}\| - \|\mathbf{z} - \mathbf{x}\|}{2\sqrt{\|\mathbf{x}\|\|\mathbf{z}\|}}, \tag{21}$$

is stationary reducible with the deformation:

$$\mathbf{\Phi}(x_1, x_2) = \Big(\ln\big(\sqrt{x_1^2 + x_2^2}\big), \arctan(x_2/x_1) \Big)^T,$$

yielding the stationary kernel:

$$K_S^*(u_1, u_2) = \cosh(u_1/2) - \sqrt{(\cosh(u_1/2) - \cos(u_2))/2}. \tag{22}$$

Effectively, it is straightforward to check with some algebra that (22) evaluated at:

$$\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{z}) = \left(\ln\left(\frac{\|\mathbf{x}\|}{\|\mathbf{z}\|}\right), \arctan(x_2/x_1) - \arctan(z_2/z_1) \right)^T,$$

yields the kernel (21). Perrin and Senoussi (1999, 2000) characterize such deformations $\mathbf{\Phi}$. Specifically, if $\mathbf{\Phi}$ and its inverse are differentiable in $\mathbb{R}^d$, and $K(\mathbf{x}, \mathbf{z})$ is continuously differentiable for $\mathbf{x} \neq \mathbf{y}$, then K satisfies (20) if and only if:

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

where $Q_{\mathbf{\Phi}}$ is the Jacobian of $\mathbf{\Phi}$ and $D_{\mathbf{x}}$ denotes the partial derivatives operator with respect to $\mathbf{x}$. It can easily be checked that the kernel (21) satisfies the above equation (23). Unfortunately, not all nonstationary kernels can be reduced to stationarity through a deformation $\mathbf{\Phi}$. Consider for instance the kernel in $\mathbb{R}^1$:

$$K(x, z) = \exp(2 - x^6 - z^6), \tag{24}$$

which is positive definite as can be seen from (6). It is obvious that $K(x, z)$ 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 $\mathbf{\Phi}$ such that:

$$K(\mathbf{x}, \mathbf{z}) = K_1\Big(\frac{\mathbf{\Phi}(\mathbf{x}) + \mathbf{\Phi}(\mathbf{z})}{2}\Big)K_2\big(\mathbf{\Phi}(\mathbf{x}) - \mathbf{\Phi}(\mathbf{z})\big), \tag{25}$$

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:

$$\mathbf{\Phi}(x) = \frac{x^3}{3} - \frac{1}{3}, \tag{26}$$

yielding:

$$K_1(u) = \exp\left(-18u^2 - 12u\right) \tag{27}$$

$$K_2(v) = \exp\left(-\frac{9}{2}v^2\right). \tag{28}$$

Here again, it can easily be checked from (27), (28), and (26) that:

$$K_1\Big(\frac{\mathbf{\Phi}(x) + \mathbf{\Phi}(z)}{2}\Big)K_2\big(\mathbf{\Phi}(x) - \mathbf{\Phi}(z)\big) = \exp(2 - x^6 - z^6).$$

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, $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^p$, cannot be reduced to stationarity or local stationarity with a bijective transformation $\mathbf{\Phi}$. Further research is needed to characterize such kernels.

6. Conclusion

In this paper, we have described several classes of kernels that can be used for machine learning: stationary (anisotropic/isotropic/compactly supported), locally stationary, nonstationary and separable nonstationary kernels. Each class has its own particular properties and spectral representation. The latter allows for the design of many new kernels in each class. We have not addressed the question of which class is best suited for a given problem, but we hope that further research will emerge from this paper. It is indeed important to find adequate classes of kernels for classification, regression, density estimation, and clustering. 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. For instance, locally stationary kernels are in fact separable kernels in a new coordinate system defined by $(\mathbf{x} + \mathbf{z})/2$ and $\mathbf{x} - \mathbf{z}$, and as already mentioned, provide computational advantages when dealing with massive data sets.

Acknowledgments

I would like to acknowledge support for this project from U.S. Army TACOM Research, Development and Engineering Center under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 634, Contract No. DAAH04-96-C-0086, TCN 00-131). I would like to thank David Gorsich from U.S. Army TACOM, Olivier Perrin, as well as the Editors and two anonymous reviewers, for their comments that improved the manuscript.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. American Mathematical Soc.*, 68: 337–404, 1950.

S. Bochner. *Harmonic Analysis and the Theory of Probability.* University of California Press, Los Angeles, California, 1955.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer, New York, 1991.

G. Christakos. On the problem of permissible covariance and variogram models. *Water Resources Research*, 20(2):251–265, 1984.

G. Christakos. *Modern Spatiotemporal Geostatistics.* Oxford University Press, New York, 2000.

G. Christakos and V. Papanicolaou. Norm-dependent covariance permissibility of weakly homogeneous spatial random fields and its consequences in spatial statistics. *Stochastic Environmental Research and Risk assessment*, 14(6):471–478, 2000.

N. Cressie. *Statistics for Spatial Data.* John Wiley & Sons, New York, 1993.

N. Cressie and H.-C. Huang. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1340, 1999.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods.* Cambridge University Press, Cambridge, 2000.

M. G. Genton and O. Perrin. On a time deformation reducing nonstationary stochastic processes to local stationarity. *Technical Report NCSU*, 2001.

J. R. Gilbert, C. Moler, and R. Schreiber. Sparse matrices in MATLAB: design and implementation. *SIAM Journal on Matrix Analysis*, 13(1):333–356, 1992.

T. Gneiting. Compactly supported correlation functions. *Journal of Multivariate Analysis*, to appear, 2002a.

T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, to appear, 2002b.

A. Graham. *Kronecker Products and Matrix Calculus: with Applications.* Ellis Horwood Limited, New York, 1981.

M. Loève. Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162, 1946.

M. Loève. *Fonctions aléatoires du second ordre.* In: Processus Stochastiques et Mouvement Brownien (P. Lévy), Gauthier-Villars, Paris, 1948.

B. B. Mandelbrot and J. W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

B. Matérn. *Spatial Variation.* Springer, New York, 1960.

G. Matheron. The intrinsic random functions and their applications. *J. Appl. Probab.*, 5: 439–468, 1973.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, A 209:415–446, 1909.

O. Perrin and R. Senoussi. Reducing non-stationary stochastic processes to stationarity by a time deformation. *Statistics and Probability Letters*, 43(4):393–397, 1999.

O. Perrin and R. Senoussi. Reducing non-stationary random fields to stationarity and isotropy using a space deformation. *Statistics and Probability Letters*, 48(1):23–32, 2000.

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(3):811–841, 1938.

R. A. Silverman. Locally stationary random processes. *IRE Transactions Information Theory*, 3:182–187, 1957.

R. A. Silverman. A matching theorem for locally stationary random processes. *Communications on Pure and Applied Mathematics*, 12:373–383, 1959.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, New York, 1999.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

V. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

A. M. Yaglom. Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory of Probability and its Applications*, 2:273–320, 1957.

A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions, Vol. I & II.* Springer Series in Statistics, New York, 1987.