ON CERTAIN GROUP INVARIANT MERCER KERNELS

Bernd-Jürgen Falkowski
University of Applied Sciences Stralsund, Department of Economics
Zur Schwedenschanze 15, D-18435 Stralsund, Germany

Key words: Mercer Kernels, Group Invariance, Radial Functions

Abstract: For the construction of support vector machines Mercer Kernels are of considerable importance. Since the conditions of Mercer's theorem are hard to verify in general, a systematic (constructive) description of Mercer kernels which are invariant under a transitive group action is provided. As an example kernels on Euclidean space invariant under the Euclidean motion group are treated. En passant a minor but confusing error in a seminal paper due to Gangolli is rectified. In addition an interesting relation to radial basis functions is exhibited.

1 INTRODUCTION

In recent years support vector machines (SVMs), cf. e.g. (Cristianini et al., 2000; Vapnik, 1998), for the theoretical background and (Shashua and Levin, 2002; Shashua and Levin, 2003; Vapnik, 1998) for practical applications, have received much attention. In this context Mercer kernels, cf. e.g. (Cristianini et al., 2000), p. 35 for Mercer's theorem, which are important building blocks of such machines, have frequently been used. These kernels determine (implicitly) the feature maps of SVMs and hence their separation capability, cf. (Cover, 1965). Thus it seems somewhat surprising that, apart from three basic types of kernels, cf. e.g. (Haykin, 1999), p. 333, and some construction rules, cf. e.g. (Cristianini et al., 2000), pp. 42-44, (Shawe-Taylor and Cristianini, 2004), p. 75-76, very little appears to be known about such kernels amongst Neural Network researchers. This is all the more surprising since the conditions of Mercer's theorem are not easily verifiable in general.

Hence it seems worthwhile to apply some mathematical results which have, in essence, been known for quite some time, to provide a complete description of all Mercer kernels which are invariant under a transitive group action. This is all the more the case since transitive group actions include the group of proper rigid motions in Euclidean space (this, of course, being of interest for practical applications, cf. (Schölkopf et al., 1999), p. 339 and p. 349). As an interesting consequence the important role of radial basis functions is seen to result from the invariance property of the kernels. En passant a minor but confusing error in (Gangolli, 1967) is rectified.

The basic mathematical results that are of interest here mainly stem from the seminal works of (Gangolli, 1967) and (Parthasarathy and Schmidt, 1972) and also from (Falkowski, 1986) (amongst the Neural Network community they seem to be little-known; even Wahba in (Wahba, 1999) does not mention them).

Being intimately connected to an abstract version of the Levy-Khinchine formula they have previously chiefly been employed to derive the structure of Mercer kernels that are described by infinitely divisible positive definite functions., cf. (Gangolli, 1967; Falkowski, 2001; Falkowski, 2003). In this paper, however, the transitivity of the group action is essential. The reader is invited to consult Minsky's corresponding results in (Minsky and Papert, 1990), (the group invariance theorem, p. 48 and theorem 2.4. p. 53) where finite groups are considered.

Note that only continuous kernels will be treated here. For details and further background information the reader is referred to (Bishop, 2006; Falkowski, 1986; Gangolli, 1967; Parthasarathy and Schmidt, 1972). For technical convenience all kernels considered here will be complex-valued in general since this greatly simplifies the technical problems concerning unitary representations of the symmetry groups.

Falkowski B. (2009).
ON CERTAIN GROUP INVARIANT MERCER KERNELS.
In *Proceedings of the International Joint Conference on Computational Intelligence*, pages 517-521
DOI: 10.5220/0002277805170521

2 BASIC FACTS

In order to be able to proceed with the relevant computations some basic definitions are needed.

2.1 Definition

Given a topological space X and a continuous function K: X×X → C (the complex numbers), K is called a *positive definite (p.d.) kernel on* X×X if it satisfies

a) $K(x,y) = \overline{K(y,x)}$ for arbitrary x, y ∈ X

b) Given $x_1, x_2, \ldots, x_n \in X$, $K(x_i,x_j)$ is p.d. as a matrix

Hence, by remark 3.7 in (Cristianini et al., 2000), p. 35, a Mercer kernel is just a real-valued positive definite kernel.

2.1.1 Example

If X is a complex Hilbert space with inner product denoted by <.. , .>, then the kernel K defined by

K(**x**,**y**):= <**x, y**> for arbitrary **x**, **y** ∈ X

is positive definite. Note though that in general the space X is not required to carry a vector space structure (although for many interesting examples it does).

2.1.2 Example

Suppose that a p.d. kernel K and a polynomial p with positive coefficients are given. Then p(K) is also a p.d. kernel.

Proof (sketch): Linear combinations of p.d. kernels with positive coefficients are obviously p.d. It remains to show that products of p.d. kernels are again p.d. This may be achieved by considering p.d. matrices **A**:= $[a_{ij}]$ and **B**:= $[b_{ij}]$ as covariance matrices of two independent normally distributed random variables **X**:= $[X_1, X_2, \ldots, X_n]$ and **Y**:= $[Y_1, Y_2, \ldots, Y_n]$ with mean vector zero. Then the matrix **C**:= $[a_{ij}*b_{ij}]$ is the covariance matrix of **Z**:= $[X_1*Y_1, X_2*Y_2, \ldots, X_n*Y_n]$ and hence p.d. Q.E.D.

2.1.3 Example

Suppose that f is the characteristic function (Fourier transform) of a probability measure on the real line, then the kernel $K(t_1,t_2):= f(t_2-t_1)$ is well-known to be positive definite and f is called a positive definite function.

For further examples and the explicit relation of kernels to feature maps the reader is referred to (Shawe-Taylor and Cristianini, 2004), pp. 47-84. Clearly, however, in this way no systematic description is obtained.

With the aim to get more explicit information the invariance condition under a transitive group action is introduced.

2.2 Definition

Let G be a topological Group with identity e and X be a topological space, as before. G is said to *act continuously* on X if

1. for every fixed g ∈ G, the map x → gx is a bijection.
2. ex = x for all x ∈ X.
3. $g_1(g_2x) = (g_1g_2)x$ for all $g_1, g_2 \in G$, x ∈ X.
4. (g, x) → gx is continuous.
5. for every fixed g ∈ G, the map x → gx is a homeomorphism of X.

The action is *transitive* if for every x, y ∈ X there exists a g ∈ G such that gx = y.

A p.d. kernel K is said to be *invariant* under G if

K(gx,gy) = K(x,y) for all g∈G.

The following theorem describes G invariant kernels in the sense of definition 2.2 in terms of unitary representations of G.

2.3 Theorem

Let X be a topological space and let G be a group acting continuously on it. Suppose that K is a p.d. kernel on X×X invariant under G.

Then there exists a complex Hilbert space H and a weakly continuous unitary representation g → U(g) of G in H (i.e. $U(g_1g_2) = U(g_1)U(g_2)$ and the map g → $<U(g)v_1,v_2>$ is continuous for every $\mathbf{v}_1, \mathbf{v}_2 \in H$) and a continuous map **v**: X → H such that the vectors **v**(x) span H and

1. K(x,y) = <**v**(x),**v**(y)>
2. **v**(gx) = U(g) **v**(x)

Proof (rough sketch): This is essentially a consequence of the Kolmogorov consistency theorem. A detailed proof is provided in (Parthasarathy and Schmidt, 1972), theorems 1.2 and 2.7 as well as remark 2.8 Q.E.D.

3 INVARIANT KERNELS

Suppose now that G acts transitively on X and choose a fixed $x_0 \in X$. Further let

$G(x_0):= \{g \in G | gx_0 = x_0\}$

be the stability subgroup of x_0 in G.

3.1 Lemma

The map $gx_0 \to gG(x_0)$ defines a bijection between X and the space of left cosets $G/G(x_0)$.
Proof: Simple computation. Q.E.D.

In addition the G-action on X corresponds to the following G-action on $G/G(x_0)$

$$g_1(gG(x_0)) := (g_1 g)G(x_0).$$

Thus an invariant kernel K on X may equally well be considered as an invariant kernel K' on $G/G(x_0)$ if the G-action is transitive. Note also that any function on $G/G(x_0)$ may be considered as a function on G that is constant on left cosets (by simply assigning the value of a left coset to all the elements in that coset) and vice versa. Thus one obtains the following theorem.

3.2 Theorem

Suppose that f is a positive definite function on G (i.e. the kernel $H(g_1,g_2) := f(g_2^{-1}g_1)$ is positive definite) that is bi-invariant under $G(x_0)$ (i.e. $f(k_1 g k_2) = f(g)$ for all $k_1,k_2 \in G(x_0)$). Then the kernel K defined by

$$K(g_1x_0,g_2x_0) := f(g_2^{-1}g_1)$$

is positive definite and invariant under G. Moreover every positive definite kernel on X invariant under G is of this form.

Proof: ($\Rightarrow$)

(i) K is well defined

Suppose $g_1x_0 = g_1'x_0$ and $g_2x_0 = g_2'x_0$. Then $g_2' = g_2k_1$ and $g_1' = g_1k_2$ for some $k_1, k_2 \in G(x_0)$. Hence $f(g_2'^{-1}g_1') = f(k_1^{-1}g_2^{-1}g_1k_2) = f(g_2^{-1}g_1)$ by bi-invariance of f under $G(x_0)$.

(ii) The positive definiteness of K follows from the positive definiteness of f.

(iii) The G- invariance of K is immediate from the definitions.

($\Leftarrow$)

From theorem 2.3 the following holds

$$\begin{aligned} K(x,y) &= K(g_1x_0,g_2x_0) \text{ for suitable } g_1, g_2 \text{ by transitivity} \\ &= \langle \mathbf{v}(g_1x_0),\mathbf{v}(g_2x_0)\rangle \\ &= \langle U(g_2^{-1}g_1)\mathbf{v}(x_0),\mathbf{v}(x_0)\rangle \\ &= f_1(g_2^{-1}g_1) \text{ say.} \end{aligned}$$

Now clearly f_1 is a positive definite function on G which is bi-invariant under $G(x_0)$. Q.E.D.

Corollary to 3.2

If $G(x_0)$ is a normal subgroup of G, then every G-invariant p.d. kernel on X is given via a p.d. function on $G/G(x_0)$.

Proof: Since $G(x_0)$ is normal $G/G(x_0)$ carries a group structure with multiplication ° given by

$$g_1G(x_0) \circ g_2G(x_0) := g_1g_2G(x_0)$$

By lemma 3.1 X may be identified with the group $G/G(x_0)$ and under this correspondence x_0 is mapped to the coset $G(x_0)$, i.e. the identity element of the group $G/G(x_0)$. So from the second part of 3.2 the following holds (denoting by $\pi : G \to G/G(x_0)$ the natural projection).

$$\begin{aligned} K(x,y) &= K'(g_1G(x_0),g_2G(x_0)) \text{ for suitable } g_1, g_2 \text{ by transitivity} \\ &= \langle \mathbf{v}(\pi(g_1)),\mathbf{v}(\pi(g_2))\rangle \\ &= \langle U(\pi(g_2^{-1})\pi(g_1))\mathbf{v}(G(x_0)),\mathbf{v}(G(x_0))\rangle \\ &= f_2(\pi(g_2)^{-1}\pi(g_1)) \text{ say.} \end{aligned}$$

Note that in the above calculation the kernel corresponding to K after identifying X with $G/G(x_0)$ has been denoted by K', and U describes a unitary representation of $G/G(x_0)$. Moreover f_2 may also be considered as a bi-invariant p.d. function on G by assigning constant values to the cosets (since left cosets are also right cosets in this case because $G(x_0)$ is assumed to be normal, there is no ambiguity).

4 EXAMPLES

In order to underline the importance of the above results some examples are provided below.

4.1 Example (G Acting on Itself)

(i) Suppose X = G and G acts on itself by left multiplication. Further let $x_0 := e$, the identity element of G. Then $G(x_0) = \{e\}$ and every p.d. kernel on G is clearly of the form

$$\begin{aligned} K(g_1,g_2) &= f(g_2^{-1}g_1) \\ &= \langle U(g_2^{-1}g_1)\mathbf{v}(e),\mathbf{v}(e)\rangle \end{aligned}$$

This is, of course, a classical result due to Gelfand and Raikov.

(ii) Even more can be said if G is locally compact second countable and abelian. Then every unitary representation U of G is the direct sum of a direct integral (for the technical details see (Naimark, 1960)) and the trivial representation. Hence there exist a measure space (Ω, S, μ) and a measurable map $\tau:\Omega \to G$ such that $\tau(\omega)$ is a nontrivial character for every ω (homomorphism into the complex unit circle) and that

$$U = \int \tau(\omega)\, d\mu(\omega) \oplus I$$

(again, for the notation and technical details the reader is referred to (Naimark, 1960).)

(iii) As a concrete version of (ii) above consider $\Omega = R^n$ (Euclidean n space) and let

$\tau(\omega)(\mathbf{x}) := e^{i\langle \mathbf{x},\omega\rangle}$,
$d\mu(\omega) := [1/(2\pi)^{n/2}] \exp[-\|\omega\|^2/2]d\omega$
$x_0(\omega) := 1$,
$\mathbf{x_0} := \int x_0(\omega)d\mu(\omega)$ (as direct integral)
and $U(\mathbf{x}) := \int e^{i\langle \mathbf{x},\omega\rangle} d\mu(\omega)$ (as direct integral)..
Then
$\langle U(\mathbf{x})\, \mathbf{x_0}, \mathbf{x_0}\rangle = \exp[-\|\mathbf{x}\|^2/2]$.
Thus from the above "abstract nonsense" the characteristic function of the standard Gaussian distribution is obtained.

4.2 Example (Euclidean Motion Group)

Here the group G of all proper rigid motions of Euclidean n space is considered. Thus let SO(n) denote the group of all proper rotations about the origin and T the group of all translations. Then obviously G acts transitively on R^n. Moreover it is well-known that every $g \in G$ can uniquely be written as
$g = tr$ where $r \in SO(n)$ and $t \in T$.
Thus the Euclidean motion group is a semi-direct product (see e.g. (Mackey, 1968), pp. 37 – 45, for further information on semi-direct products) of SO(n) and R^n as follows. Suppose that the rotation $r_{(i)}$ is represented by a matrix $\mathbf{A_{(i)}}$ and the translation $t_{(i)}$ is represented by a vector $\mathbf{b_{(i)}}$ then if $g_i = t_i r_i$
$g_1g_2\mathbf{x} = g_1(\mathbf{A_2x} + \mathbf{b}_2)$
$= \mathbf{A_1A_2x} + \mathbf{b_1} + \mathbf{A_1b_2}$
Hence if an element of G is now denoted by $(\mathbf{b}, \mathbf{A})$, the following group operation is obtained
$(\mathbf{b_1}, \mathbf{A_1}) \circ (\mathbf{b_2}, \mathbf{A_2}) = (\mathbf{b_1} + \mathbf{A_1b_2}, \mathbf{A_1A_2})$.
Clearly R^n appears as a normal subgroup here (the first component of the Cartesian product).

Thus if the point x_0 is taken to be the origin then the stability subgroup is simply SO(n) and the quotient is R^n.

Incidentally, in (Gangolli, 1967) it is mistakenly claimed that SO(n) is a normal subgroup. This is not true and thus the quotient does not carry a group structure (addition) but must be considered as space of cosets. However, this is only confusing since the results given in (Gangolli, 1967) are not affected by the error.

Now theorem 3.2 (albeit because of the above remark not the corollary) can be applied to obtain

4.2.1 Theorem

Every p.d. kernel K on R^n that is invariant under the Euclidean motion group G is given by a p.d. radial function on R^n, i.e. a function that depends only on the Euclidean distance from the origin.

Proof: Since R^n considered as quotient does not carry a group structure, it has to be treated as the space of left cosets. Nevertheless, by theorem 3.2 the kernel K is still described by a bi-invariant p.d. function on G which must, a fortiori, be also p.d. on R^n. Starting with an arbitrary p.d. function f_2 on R^n it can be extended to G by defining it to be constant on left cosets. However, this will not suffice to make it bi-invariant. Indeed since (denoting the extended function by f_1)
$f_1((\mathbf{b_1}, \mathbf{A_1}) \circ (\mathbf{0}, \mathbf{A_2})) = f_1((\mathbf{b_1}, \mathbf{A_1A_2})) =$
$f_1((\mathbf{b_1}, \mathbf{I})) = f_1((\mathbf{b_1}, \mathbf{A_1})) = f_2(\mathbf{b_1})$
and thus f_1 is right invariant.
However, since
$f_1((\mathbf{0}, \mathbf{A_1}) \circ (\mathbf{b_2}, \mathbf{A_2})) =$
$f_1((\mathbf{A_1b_2}, \mathbf{A_1A_2}))$
one must demand for left invariance that for arbitrary rotations $\mathbf{A_1}$ $f_1((\mathbf{A_1b_2}, \mathbf{A_2})) = f_1((\mathbf{b_2}, \mathbf{A_2}))$ and this in turn leads to the requirement that f_2 must be a p.d. radial function on R^n, i.e. $f_2(\mathbf{b}) = h(\|\mathbf{b}\|)$, where h is some other function and $\|.\|$ denotes the length of a vector in R^n. Q.E.D.

4.3 Remarks

(i) It is interesting to observe that applying the invariance condition in the case of the Euclidean motion group leads to a radial basis function kernel and thus to RBF networks. For example if one considers the Gaussian kernel
$K(\mathbf{x},\mathbf{y}) := \exp[-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2]$,
Then the feature map ϕ represents the elements of the feature space as functions in a Hilbert space by
$\phi(\mathbf{x}) = K(\mathbf{x}, .)$,
where the scalar product between functions is then given by
$\langle \Sigma_i\alpha_i K(\mathbf{x}_i, .), \Sigma_j\beta_j K(\mathbf{x}_j, .)\rangle =$
$\Sigma_i\Sigma_j\alpha_i\beta_j K(x_i,x_j)$.
For further details see also (Shawe-Taylor and Cristianini, 2004), p. 77 and p. 297.
It should also be observed that in the case of finite permutation groups as treated by Minsky, cf. (Minsky and Papert, 1990), p. 53, the invariance condition leads to very severe restrictions.
(ii) It can be shown that any continuous function on a compact interval can be approximated with arbitrary accuracy by a linear combination of RBFs, cf. (Bishop, 2006), p. 299, (Powell, 1987).
(iii) From example 4.1 (ii) and theorem 4.2.1 it is now possible to obtain a complete description of all p.d. kernels on R^n that are invariant under the Euclidean motion group. Of course, it must be admitted that here only complex valued kernels have been considered because of technical convenience whilst

generally the real valued ones will be of major interest. Moreover there is the question of choosing a suitable measure for the direct integrals. Nevertheless, modulo these complications the explicit description is arrived at by "radializing" the positive definite functions on R^n along the lines described in (Gangolli, 1967), p. 134 for Levy Schoenberg Kernels. For further explicit examples see (Gangolli, 1967; Falkowski, 2001; Falkowski, 2003). In particular in (Gangolli, 1967) several real-valued Mercer kernels are explicitly described.

5 CONCLUSIONS

Some results from pure mathematics have been employed to derive a detailed description of group invariant Mercer kernels, where the group action was assumed to be transitive. As an application a classical theorem due to Gelfand and Raikov was recovered. Thereafter kernels invariant under the Euclidean motion group were considered in detail. A complete description (modulo some technical details) was provided. Moreover it was shown that these kernels are derived from radial functions on R^n. En passant a minor but confusing error in (Gangolli, 1967) was rectified. The connection to radial basis function networks was explained. It seems rather satisfying that using only invariance conditions (which have also very successfully been employed in an entirely different context such as quantum mechanics, cf. (Mackey, 1968) Mackey) on the kernels such explicit results can be derived for interesting practical applications, cf. (Schölkopf et al., 1999). The author is tempted to paraphrase part of Minsky and Papert's remark in (Minsky and Papert, 1990), p. 241: These methods brought the feeling of "real mathematics". ... This is still sufficiently rare in computer science to be significant. We are convinced that respect for "real mathematics" is a powerful heuristic principle, though it must be tempered with practical judgment.

REFERENCES

Bishop,C.M.: Pattern Recognition and Machine Learning Springer, (2006)

Cover, T.M.: Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. IEEE Trans. on Electronic Computers, Vol. 14, (1965), pp. 326-334

Cristianini, N.; Shawe-Taylor, J.: An Introduction to Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press, (2000)

Falkowski, B.-J.: Levy-Schoenberg Kernels on Riemannian Symmetric Spaces of Non-Compact Type. In: Probability Measures on Groups, Ed. H. Heyer, Springer Lecture Notes in Mathematics, Vol. 1210, (1986), pp. 58-67

Falkowski, B.-J.: Mercer Kernels and 1-Cohomology. In: Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, Proceedings of KES 2001, Eds. N. Bab; L.C. Jain; R.J. Howlett, IOS Press, (2001), pp. 461-465

Falkowski B.-J.: Mercer Kernels and 1-Cohomology of Certain Semi-simple Lie Groups. In: Proc. of the 7th Intl. Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003), Eds. V. Palade, R.J. Howlett, L.C. Jain, LNAI Vol. 2773, Springer Verlag, (2003)

Gangolli, R.: Positive Definite Kernels on Homogeneous Spaces. In: Ann. Inst. H. Poincare B, Vol. 3, (1967), pp. 121-225

Haykin, S.: Neural Networks, a Comprehensive Foundation. Prentice-Hall, (1999)

Mackey, G.W.: Induced Representations of Groups and Quantum Mechanics. Benjamin, (1968)

Minsky, M.L.; Papert, S.: Perceptrons. MIT Press, (Expanded Edition 1990)

Naimark, M.A. Normed Rings (English Translation). Noordhoff, Amsterdam, (1960)

Parthasarathy, K.R.; Schmidt, K.: Positive Definite Kernels, Continuous Tensor Products, and Central Limit Theorems of Probability Theory. Springer Lecture Notes in Mathematics, Vol. 272, (1972)

Powell, M.J.D.: Radial Basis Functions for Multivariable Interpolation: a Review. In J.C. Cox and M.G. Cox (Eds.) Algorithms for Approximation, Oxford University Press, (1987), pp. 143-167

Schölkopf, B.; Smola, A.J.; Müller, K.-R.: Kernel Principal Component Analysis. In: Schölkopf, B.; Burges, J.C.; Smola, A.J. (Eds.): Advances in Kernel Methods, Support Vector Learning. MIT Press, (1999), pp. 327-352

Shashua, A.; Levin, A.: Taxonomy of Large Margin Principle Algorithms for Ordinal Regression Problems. Technical Report 2002-39, Leibniz Center for Research, School of Computer Science and Eng., the Hebrew University of Jerusalem, (2002)

Shashua, A.; Levin, A.:Ranking with Large Margin Principle: Two Approaches, NIPS 14, (2003)

Shawe-Taylor, J.; Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, (2004)

Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, (1998)

Wahba, G.: Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV. In: Schölkopf, B.; Burges, J.C.; Smola, A.J. (Eds.): Advances in Kernel Methods, Support Vector Learning. MIT Press, (1999), pp. 69-88