Covariance Functions

by That Guy Who Wrote The Book That was Way Too Focused On Machine Learning to be Of Use and Incorrected Stated Complex Exponentials as being Eigenfunctions of statiomary kernels rather than covariance operators having stationary kernels because it doesnt even make sense to say a function has an eigenfunction

Table of contents

1 Covariance Functions

It has been seen that a covariance function is the crucial ingredient in a Gaussian process predictor, as it encodes the assumptions about the function which is to be learned. From a slightly different viewpoint it is clear that in supervised learning the notion of similarity between data points is crucial; it is a basic assumption that points with inputs x which are close are likely to have similar target values y, and thus training points that are near to a test point should be informative about the prediction at that point. Under the Gaussian process view it is the covariance function that defines nearness or similarity.

An arbitrary function of input pairs x and x' will not, in general, be a valid covariance function.[1] The purpose of this chapter is to give examples of some commonly-used covariance functions and to examine their properties. Section 1 defines a number of basic terms relating to covariance functions. Section 2 gives examples of stationary, dot-product, and other non-stationary covariance functions, and also gives some ways to make new ones from old. Section 3 introduces the important topic of eigenfunction analysis of covariance functions, and states Mercer's theorem which allows the expression of the covariance function (under certain conditions) in terms of its eigenfunctions and eigenvalues. The covariance functions given in section 2 are valid when the input domain $\mathcal{X}$ is a subset of $\mathbb{R}^D$. In section 4 methods are described to define covariance functions when the input domain is over structured objects such as strings and trees.

1 Preliminaries

A stationary covariance function is a function of $x - x'$. Thus it is invariant to translations in the input space.[2] For example the squared exponential covariance function given in equation (9) is stationary. If further the covariance function is a function only of $|x - x'|$ then it is called isotropic; it is thus invariant to all rigid motions. For example the squared exponential covariance function given in equation (9) is isotropic. As k is now only a function of $r = |x - x'|$ these are also known as radial basis functions (RBFs).

If a covariance function depends only on x and x' through $x \cdot x'$ it is called a dot product covariance function. A simple example is the covariance function $k(x, x') = \sigma_0^2 + x \cdot x'$ which can be obtained from linear regression by putting $\mathcal{N}(0, 1)$ priors on the coefficients of x_d $(d = 1, \ldots, D)$ and a prior of $\mathcal{N}(0, \sigma_0^2)$ on the bias (or constant function) 1, see equation (15). Another important example is the inhomogeneous polynomial kernel $k(x, x') = (\sigma_0^2 + x \cdot x')^p$ where p is a positive integer. Dot product covariance functions are invariant to a rotation of the coordinates about the origin, but not translations.

A general name for a function k of two arguments mapping a pair of inputs $x \in \mathcal{X}$, $x' \in \mathcal{X}$ into $\mathbb{R}$ is a kernel. This term arises in the theory of integral operators, where the operator T_k is defined as

$$(T_k f)(x) = \int_{\mathcal{X}} k(x, x')\, f(x')\, d\,\mu(x'), \tag{1}$$

where μ denotes a measure; see Section 3 for further discussion.

A real kernel is said to be symmetric if $k(x, x') = k(x', x)$; clearly covariance functions must be symmetric from the definition.

1. To be a valid covariance function it must be positive semidefinite, see equation (2).

2. In stochastic process theory a process which has constant mean and whose covariance function is invariant to translations is called weakly stationary. A process is strictly stationary if all of its finite dimensional distributions are invariant to translations [papoulis1991].

Given a set of input points $\{x_i | i = 1, \ldots, n\}$ the Gram matrix K can be computed whose entries are $K_{ij} = k(x_i, x_j)$. If k is a covariance function the matrix K is called the covariance matrix.

A real $n \times n$ matrix K which satisfies $Q(v) = v^\top K v \geq 0$ for all vectors $v \in \mathbb{R}^n$ is called positive semidefinite (PSD). If $Q(v) = 0$ only when $v = 0$ the matrix is positive definite. $Q(v)$ is called a quadratic form. A symmetric matrix is PSD if and only if all of its eigenvalues are non-negative. A Gram matrix corresponding to a general kernel function need not be PSD, but the Gram matrix corresponding to a covariance function is PSD.

A kernel is said to be positive semidefinite if

$$\int k(x, x')\, f(x)\, f(x')\, d\,\mu(x)\, d\,\mu(x') \geq 0 \tag{2}$$

for all $f \in L^2(\mathcal{X}, \mu)$. Equivalently a kernel function which gives rise to PSD Gram matrices for any choice of n and $\mathcal{D}$ is positive semidefinite.

For a one-dimensional Gaussian process one way to understand the characteristic length-scale of the process (if this exists) is in terms of the number of upcrossings of a level u. Adler [adler1981] states that the expected number of upcrossings $E[N_u]$ of the level u on the unit interval by a zero-mean, stationary, almost surely continuous Gaussian process is given by

$$E[N_u] = \frac{1}{2\,\pi} \sqrt{\frac{-k''(0)}{k(0)}} \exp\left(-\frac{u^2}{2\,k(0)}\right) \tag{3}$$

If $k''(0)$ does not exist (so that the process is not mean square differentiable) then if such a process has a zero at x_0 then it will almost surely have an infinite number of zeros in the arbitrarily small interval $(x_0, x_0 + \delta)$ [blake1973].

1.1 Mean Square Continuity and Differentiability

Mean square continuity and differentiability of stochastic processes is now described, following Adler [adler1981]. Let $x_1, x_2, \ldots$ be a sequence of points and x^* be a fixed point in $\mathbb{R}^D$ such that $|x_k - x^*| \to 0$ as $k \to \infty$. Then a process $f(x)$ is continuous in mean square at x^* if $E[|f(x_k) - f(x^*)|^2] \to 0$ as $k \to \infty$. If this holds for all $x^* \in A$ where A is a subset of $\mathbb{R}^D$ then $f(x)$ is said to be continuous in mean square (MS) over A. A random field is continuous in mean square at x^* if and only if its covariance function $k(x, x')$ is continuous at the point $x = x' = x^*$. For stationary covariance functions this reduces to checking continuity at $k(0)$. Note that MS continuity does not necessarily imply sample function continuity; for a discussion of sample function continuity and differentiability see Adler [adler1981].

The mean square derivative of $f(x)$ in the ith direction is defined as

$$\frac{\partial\, f(x)}{\partial\, x_i} = \lim_{h \to 0} \frac{f\,(x + h\, e_i) - f(x)}{h} \tag{4}$$

when the limit exists, where the limit is in mean square and e_i is the unit vector in the ith direction. The covariance function of $\partial f(x)/\partial x_i$ is given by $\frac{\partial^2 k(x,x')}{\partial x_i \partial x_i'}$. These definitions can be extended to higher order derivatives. For stationary processes, if the $2k$th-order partial derivative $\frac{\partial^{2k} k(x)}{\partial^2 x_{i_1} \ldots \partial^2 x_{i_k}}$ exists and is finite at $x = 0$ then the kth order partial derivative $\frac{\partial^k f(x)}{\partial x_{i_1} \ldots x_{i_k}}$ exists for all $x \in \mathbb{R}^D$ as a mean square limit. Notice that it is the properties of the kernel k around 0 that determine the smoothness properties (MS differentiability) of a stationary process.

2 Examples of Covariance Functions

In this section covariance functions are considered where the input domain $\mathcal{X}$ is a subset of the vector space $\mathbb{R}^D$. More general input spaces are considered in Section 4. The section starts in Section 2.1 with stationary covariance functions, then considers dot-product covariance functions in Section 2.2 and other varieties of non-stationary covariance functions in Section 2.3. An overview of some commonly used covariance functions is given in Table 1 and in Section ? general methods for constructing new kernels from old are described. There exist several other good overviews of covariance functions, see e.g. Abrahamsen [abrahamsen1997].

2.1 Stationary Covariance Functions

In this section (and Section 3) it will be convenient to allow kernels to be a map from $x \in \mathcal{X}, x' \in \mathcal{X}$ into $\mathbb{C}$ (rather than $\mathbb{R}$). If a zero-mean process f is complex-valued, then the covariance function is defined as

$$k(x, x') = E\left[f(x)\, f^*(x')\right] \tag{5}$$

where $*$ denotes complex conjugation. A stationary covariance function is a function of $\tau = x - x'$. Sometimes in this case k will be written as a function of a single argument, i.e. $k(\tau)$.

The covariance function of a stationary process can be represented as the Fourier transform of a positive finite measure.

Theorem 1. *[Bochner's theorem] A complex-valued function k on $\mathbb{R}^D$ is the covariance function of a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^D$ if and only if it can be represented as*

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i \boldsymbol{s}\cdot\tau} d\,\mu(\boldsymbol{s}) \tag{6}$$

where μ is a positive finite measure.

If μ has a density $S(\boldsymbol{s})$ then S is called the spectral density or power spectrum corresponding to k.

In the case that the spectral density $S(\boldsymbol{s})$ exists, the covariance function and the spectral density are Fourier duals of each other as shown in

$$k(\tau) = \int S(\boldsymbol{s})\, e^{2\pi i \boldsymbol{s}\cdot\tau}\, d\,\boldsymbol{s} \tag{7}$$
$$S(\boldsymbol{s}) = \int k(\tau)\, e^{-2\pi i \boldsymbol{s}\cdot\tau}\, d\,\tau$$

This is the Wiener-Khintchine theorem [chatfield1989]; see Appendix A.8 for details of Fourier transforms.

Notice that the variance of the process is $k(0) = \int S(\boldsymbol{s})\, d\,\boldsymbol{s}$ so the power spectrum must be integrable to define a valid Gaussian process.

The complex exponentials $e^{2\pi i \boldsymbol{s}\cdot x}$ are eigenfunctions of a stationary kernel with respect to Lebesgue measure (see Section 3 for further details).

If the covariance function is isotropic (so that it is a function of r, where $r = |\tau|$) then $S(\boldsymbol{s})$ is a function of $|\boldsymbol{s}|$ only [adler1981]. In this case the integrals in (7) can be simplified by changing to spherical polar coordinates and integrating out the angular variables, yielding a Hankel transform:

$$k(r) = \frac{2\,\pi}{r^{D/2-1}} \int_0^\infty S(s)\, J_{D/2-1}(2\,\pi\, r\, s)\, s^{D/2}\, d\, s \tag{8}$$
$$S(s) = \frac{2\,\pi}{s^{D/2-1}} \int_0^\infty k(r)\, J_{D/2-1}(2\,\pi\, r\, s)\, r^{D/2}\, d\, r$$

where $J_{D/2-1}$ is a Bessel function of order $D/2-1$.

Common Stationary Covariance Functions:

Below are several commonly-used covariance functions as defined in Table 1. All expressions refer to covariance $k(r) = k(|x - x'|)$ unless otherwise indicated.

1. **Squared Exponential (SE):**

$$k_{\mathrm{SE}}(r) = \exp\left(-\frac{r^2}{2\,\ell^2}\right) \tag{9}$$

   Parameter ℓ is the characteristic length-scale. This kernel is infinitely differentiable. The SE spectral density is $S(s) = (2\,\pi\,\ell^2)^{D/2} \exp(-2\,\pi^2\,\ell^2\,s^2)$.

2. **Matérn class:**

$$k_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\,\nu}\, r}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\,\nu}\, r}{\ell}\right) \tag{10}$$

   with positive parameters ν, ℓ, and K_ν a modified Bessel function [abramowitz1965]. See Section 2 for special cases.

3. **Exponential:**

$$k(r) = \exp\left(-r/\ell\right) \tag{11}$$

A special case of the Matérn with $\nu = 1/2$. Not MS differentiable. Related to the Ornstein-Uhlenbeck process.

4. **γ-Exponential:**

$$k(r) = \exp\left(-\left(\frac{r}{\ell}\right)^{\gamma}\right), \quad 0 < \gamma \leq 2 \tag{12}$$

5. **Rational Quadratic:**

$$k_{\mathrm{RQ}}(r) = \left(1 + \frac{r^2}{2\,\alpha\,\ell^2}\right)^{-\alpha} \tag{13}$$

This is a scale mixture of SE kernels with Gamma-distributed length-scales.

6. **Piecewise polynomial with compact support:**

$$\begin{aligned}
k_{pp}^{D,0}(r) &= (1-r)_+^j \\
k_{pp}^{D,1}(r) &= (1-r)_+^{j+1}\left((j+1)\,r+1\right) \\
k_{pp}^{D,2}(r) &= (1-r)_+^{j+2}\left((j^2+4\,j+3)\,r^2+(3\,j+6)\,r+3\right)/3 \\
k_{pp}^{D,3}(r) &= (1-r)_+^{j+3}\left((j^3+9\,j^2+23\,j+15)\,r^3+(6\,j^2+36\,j+45)\,r^2+(15\,j+\right. \\
&\qquad \left.45)\,r+15\right)/15
\end{aligned} \tag{14}$$

where $j = \lfloor D/2 \rfloor + q + 1$.

See [wendland2005] for conditions of positive-definiteness in different input dimensions.

covariance function	expression	Stationary? (S)	Nondegenerate? (ND)
constant	σ_0^2	✓	
linear	$\sum_{d=1}^{D} \sigma_d^2\, x_d\, x'_d$		
polynomial	$(x \cdot x' + \sigma_0^2)^p$		
squared exponential	$\exp\left(-r^2/2\,\ell^2\right)$	✓	✓
Matérn	see Eq. (10)	✓	✓
exponential	$\exp\left(-r/\ell\right)$	✓	✓
γ-exponential	$\exp\left(-(r/\ell)^{\gamma}\right)$	✓	✓
rational quadratic	$(1 + r^2/(2\,\alpha\,\ell^2))^{-\alpha}$	✓	✓
neural network	see Eq. (18)		✓

Table 1. Summary of several commonly-used covariance functions. $r = |x - x'|$. See Section 2 for details.

2.2 Dot Product Covariance Functions

The kernel

$$k(x, x') = \sigma_0^2 + x \cdot x' \tag{15}$$

arises from Bayesian linear regression. Homogeneous if $\sigma_0^2 = 0$, inhomogeneous otherwise. It is also valid to apply a general positive-semidefinite Σ_p:

$$k(x, x') = \sigma_0^2 + x^\top \Sigma_p \, x' \tag{16}$$

Polynomials are generated via

$$k(x, x') = (\sigma_0^2 + x^\top \Sigma_p \, x')^p \tag{17}$$

for p a positive integer. See Section 2.2 for feature space construction.

2.3 Other Non-stationary Covariance Functions

Covariance functions derived from neural networks [neal1996, williams1998], convolution constructions, and kernels defined on nonlinear input transformations $u(x)$ (*warping*, see [sampson1992]) are important examples. The neural network (NN) kernel (with erf nonlinearity) is

$$k_{\mathrm{NN}}(x, x') = \frac{2}{\pi} \sin^{-1}\left(\frac{2 \, \tilde{x}^\top \Sigma \, \tilde{x}'}{\sqrt{(1 + 2 \, \tilde{x}^\top \Sigma \, \tilde{x}) \, (1 + 2 \, \tilde{x}'^\top \Sigma \, \tilde{x}')}} \right) \tag{18}$$

where $\tilde{x} = (1, x_1, ..., x_D)^\top$. More non-stationary constructions, including the Gibbs kernel, are discussed in Section 2.3.

3 Eigenfunction Analysis of Kernels

A function $\phi(\cdot)$ is an eigenfunction of kernel k (with respect to measure μ) with eigenvalue λ if

$$\int k(x, x') \, \phi(x) \; d\,\mu(x) = \lambda \, \phi(x') \tag{19}$$

holds. The ordering is such that $\lambda_1 \geq \lambda_2 \geq ...$ and eigenfunctions are orthonormal.

Theorem 2. *[Mercer's theorem] Let $(\mathcal{X}, \mu)$ be a finite measure space and $k \in L^\infty(\mathcal{X}^2, \mu^2)$ a kernel such that $T_k \colon L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ is positive definite (see Eq. (2)). Let ϕ_i be normalized eigenfunctions associated to eigenvalues $\lambda_i > 0$. Then*

1. *The sequence $\{\lambda_i\}$ is absolutely summable;*
2. $$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \, \phi_i(x) \, \phi_i^*(x') \tag{20}$$

holds μ^2 almost everywhere, with absolute and uniform convergence.

A *degenerate* kernel has finitely many nonzero eigenvalues.

For stationary kernels on $\mathbb{R}^D$, Bochner's theorem yields

$$k\,(x - x') = \int_{\mathbb{R}^D} e^{2\pi i \boldsymbol{s} \cdot (x - x')} \, d\,\mu(\boldsymbol{s}) \tag{21}$$

where the Fourier exponentials $e^{2\pi i \boldsymbol{s} \cdot x}$ are eigenfunctions.

For the SE kernel with Gaussian measure, the eigenvalues and eigenfunctions are [zhu1998]

$$\lambda_k = \sqrt{\frac{2\,a}{A}} \, B^k \tag{22}$$

$$\phi_k(x) = \exp\left(-(c - a)\, x^2\right) H_k\left(\sqrt{2\,c}\, x\right) \tag{23}$$

where H_k is the kth Hermite polynomial, $a^{-1} = 4\,\sigma^2$, $b^{-1} = 2\,\ell^2$, $c = \sqrt{a^2 + 2\,a\,b}$, $A = a + b + c$, $B = b \,/\, A$.

4 Kernels for Non-vectorial Inputs

For data such as strings, trees, or graphs, kernels are defined as follows.

4.1 String Kernels

Let $\mathcal{A}$ be a finite alphabet. The string kernel is

$$k(x, x') = \sum_{s \in \mathcal{A}^*} w_s \, \phi_s(x) \, \phi_s(x') \tag{24}$$

where $\phi_s(x)$ is the count of substring s in x, and $w_s \geq 0$. Special cases include the bag-of-characters, bag-of-words, and k-spectrum kernels.

4.2 Fisher Kernels

Given a probabilistic model $p(x|\,\theta)$, the Fisher kernel uses the score $\phi_\theta(x) = \nabla_\theta \log p(x|\theta)$:

$$k(x, x') = \phi_\theta(x)^\top M^{-1}\, \phi_\theta(x') \tag{25}$$

Usually M is the Fisher information matrix: $M = E_x\left[\phi_\theta(x)\, \phi_\theta(x)^\top\right]$ [amari1985].

5 Exercises

1. The OU process with covariance $k\,(x - x') = \exp\,(-|x - x'|\,/\,\ell)$ is the unique stationary first-order Markovian Gaussian process. See computation in Section 2.

2. Write code to draw samples from the neural network covariance function, Eq. (18) in 1-d and 2-d. Compare $\mathrm{var}(u_0)$ values.

3. Show that the random process $f(x) = \mathrm{erf}(u_0 + \sum_{i=1}^{D} u_j\, x_j)$ with $u \sim \mathcal{N}(0, \Sigma)$ has the same covariance as the NN kernel.

4. Derive Gibbs' non-stationary covariance function, Eq. (?).

5. Draw samples from Gibbs' non-stationary covariance, Eq. (?), for various $\ell(x)$.

6. Show the SE process is infinitely MS differentiable, but the OU process is not.

7. Prove eigenfunctions of a symmetric kernel are orthogonal w.r.t. μ.

8. Show the transformation $k(x, x') \mapsto \tilde{k}(x, x') = p^{1/2}(x)\, k(x, x')\, p^{1/2}(x')$ gives the same eigenvalues as with weighted measure.

9. Apply this to the SE kernel and Gaussian density. Confirm Eqs. (22) and (23).

10. Compare the SE eigenfunctions exactly to numerics via the Nyström method.

11. Let $x \sim \mathcal{N}(\mu, \sigma^2\, I)$. Show the Fisher kernel for $\mu = 0$ is the linear kernel: $k(x, x') = \frac{1}{\sigma^2}\, x \cdot x'$.

12. For a $k-1$ order Markov model string kernel, derive the Fisher score and relate it to the k-spectrum kernel.

A.6 References

Bibliography

[abrahamsen1997] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Oslo, 1997.

[abramowitz1965] M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions. Dover, New York, 1965.

[adler1981] R. J. Adler. The Geometry of Random Fields. John Wiley & Sons, New York, 1981.

[amari1985] S. Amari. Differential-Geometrical Methods in Statistics, volume 28 of Lecture Notes in Statistics. Springer-Verlag, New York, 1985.

[bach2002] F. R. Bach and M. I. Jordan. Kernel independent component analysis. Journal of Machine Learning Research, 3:1–48, 2002.

[baker1977] C. T. H. Baker. The Numerical Treatment of Integral Equations. Oxford University Press, Oxford, 1977.

[blake1973] I. F. Blake and W. C. Lindsey. Level-crossing problems for random processes. IEEE Transactions on Information Theory, 19(3):295–315, 1973.

[bracewell1986] R. N. Bracewell. The Fourier Transform and its Applications. McGraw-Hill, second edition, 1986.

[chatfield1989] C. Chatfield. The Analysis of Time Series: An Introduction. Chapman and Hall, London, fourth edition, 1989.

[collins2002] M. Collins and N. Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 263–270, 2002.

[cornford2002] D. Cornford, I. T. Nabney, and C. K. I. Williams. Adding constrained discontinuities to Gaussian process models of wind fields. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, pages 861–867. MIT Press, Cambridge, MA, 2002.

[gibbs1997] M. N. Gibbs. Bayesian Gaussian Processes for Regression and Classification. PhD thesis, University of Cambridge, 1997.

[gihman1974] I. I. Gihman and A. V. Skorohod. The Theory of Stochastic Processes I. Springer-Verlag, New York, 1974.

[gradshteyn1980] I. S. Gradshteyn and I. M. Ryzhik. Table of Integrals, Series, and Products. Academic Press, New York, 1980.

[grimmett1992] G. R. Grimmett and D. R. Stirzaker. Probability and Random Processes. Oxford University Press, Oxford, second edition, 1992.

[hand2001] D. Hand, H. Mannila, and P. Smyth. Principles of Data Mining. MIT Press, Cambridge, MA, 2001.

[hastie1990] T. J. Hastie and R. J. Tibshirani. Generalized Additive Models. Chapman and Hall, London, 1990.

[haussler1999] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California at Santa Cruz, 1999.

[hawkins1989] D. M. Hawkins. Using U statistics to derive the asymptotic distribution of Fisher's Z statistic. The American Statistician, 43(4):235–237, 1989.

[hornik1993] K. Hornik. Some new results on neural network approximation. Neural Networks, 6:1069–1072, 1993.

[jaakkola2000] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. Journal of Computational Biology, 7(1/2):95–114, 2000.

[konig1986] H. König. Eigenvalue Distribution of Compact Operators. Birkhäuser Verlag, Basel, 1986.

[leslie2003] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. Bioinformatics, 20(4):467–476, 2004.

[lodhi2001] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. Journal of Machine Learning Research, 2:419–444, 2002.

[mackay1998] D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, Neural Networks and Machine Learning, pages 133–165. Springer-Verlag, Berlin, 1998.

[matern1960] B. Matérn. Spatial Variation. Meddelanden från Statens Skogsforskningsinstitut, volume 49, number 5. Springer-Verlag, New York, second edition, 1986. [First edition published in 1960 by Statens Skogsforskningsinstitut].

[neal1996] R. M. Neal. Bayesian Learning for Neural Networks. Springer-Verlag, New York, 1996.

[ohagan1978] A. O'Hagan. Curve fitting and optimal design for prediction. Journal of the Royal Statistical Society B, 40:1–42, 1978.

[paciorek2004] C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for Gaussian process regression. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004.

[papoulis1991] A. Papoulis. Probability, Random Variables, and Stochastic Processes. McGraw-Hill, New York, third edition, 1991.

[plate1999] T. A. Plate. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models. Behaviormetrika, 26(1):29–50, 1999.

[poggio1990] T. Poggio and F. Girosi. Networks for approximation and learning. Proceedings of the IEEE, 78(9):1481–1497, 1990.

[press1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, second edition, 1992.

[ritter1995] K. Ritter, G. W. Wasilkowski, and H. Woźniakowski. Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. The Annals of Applied Probability, 5(2):518–540, 1995.

[salton1988] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.

[sampson1992] P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association, 87(417):108–119, 1992.

[saunders2003] C. Saunders, H. Tschach, and J. Shawe-Taylor. Syllables and other string kernel extensions. In Proceedings of the Nineteenth International Conference on Machine Learning, pages 530–537. Morgan Kaufmann, 2002.

[schoenberg1938] I. J. Schoenberg. Metric spaces and completely monotone functions. Annals of Mathematics, 39(4):811–841, 1938.

[scholkopf1998] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299–1319, 1998.

[scholkopf2002] B. Schölkopf and A. J. Smola. Learning with Kernels. MIT Press, Cambridge, MA, 2002.

[shawe2003] J. Shawe-Taylor and C. K. I. Williams. The stability of kernel principal components analysis and its relation to the process eigenspectrum. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems 15, pages 383–390. MIT Press, Cambridge, MA, 2003.

[stein1999] M. L. Stein. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York, 1999.

[stitson1999] M. O. Stitson, J. A. E. Weston, A. Gammerman, V. Vovk, and V. Vapnik. Theory of support vector machines. Technical Report CSD-TR-96-17, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1996.

[tsuda2002] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller. A new discriminative kernel from probabilistic models. Neural Computation, 14(10):2397–2414, 2002.

[uhlenbeck1930] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. Physical Review, 36(5):823–841, 1930.

[vishwanathan2003] S. V. N. Vishwanathan and A. J. Smola. Fast kernels for string and tree matching. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems 15, pages 585–592. MIT Press, Cambridge, MA, 2003.

[vivarelli1999] F. Vivarelli and C. K. I. Williams. Discovering hidden features with Gaussian processes regression. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, Advances in Neural Information Processing Systems 11, pages 613–619. MIT Press, Cambridge, MA, 1999.

[wahba1990] G. Wahba. Spline Models for Observational Data. SIAM, Philadelphia, 1990.

[watkins1999] C. Watkins. Dynamic alignment kernels. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers, pages 39–50. MIT Press, Cambridge, MA, 2000.

[watkins2000] C. Watkins. Dynamic alignment kernels. Technical Report CSD-TR-98-11, Department of Computer Science, Royal Holloway, University of London, 1999.

[wendland2005] H. Wendland. Scattered Data Approximation. Cambridge University Press, Cambridge, 2005.

[widom1963] H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations. Transactions of the American Mathematical Society, 109(2):278–295, 1963.

[widom1964] H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations II. Archive for Rational Mechanics and Analysis, 17(3):215–229, 1964.

[williams1998] C. K. I. Williams. Computation with infinite neural networks. Neural Computation, 10(5):1203–1216, 1998.

[yaglom1987] A. M. Yaglom. Correlation Theory of Stationary and Related Random Functions. Springer-Verlag, New York, 1987.

[zhu1998] H. Zhu, C. K. I. Williams, R. J. Rohwer, and M. Morciniec. Gaussian regression and optimal finite dimensional linear models. In C. M. Bishop, editor, Neural Networks and Machine Learning, pages 167–184. Springer-Verlag, Berlin, 1998.