MATHEMATICAL RESULTS CONCERNING KERNEL TECHNIQUES

Robert Schaback *

* *Universität Göttingen*
Institut für Numerische und Angewandte Mathematik
Lotzestraße 16–18
D–37083 Göttingen
schaback@math.uni-goettingen.de

Abstract: Black–box models based on kernels K are written as mappings of the form

$$F_\alpha(x) = \sum_j \alpha_j K(x_j, x)$$

that are intended to reproduce observational input/output data pairs (x_j, y_j) in the sense $F(x_j) \approx y_j$. Such functions have been studied in a general mathematical context for quite some time, and this contribution reviews part of the known facts and provides links to a subset of the background literature. Special emphasis is given to questions of optimality and complexity within the context of black–box modelling.

Keywords: Kernels. radial basis functions, reduction, optimality, black–box modelling.

1. MODELS

The goal of this contribution within SYSID 2003 is to make old and new mathematical results on kernels available to the system identification community, in particular within the context of black–box modelling (Sjoberg *et al.*, 1995). For easy alignment with the mathematical background, a black–box model is simplified here to be a parametrized nonlinear multivariate function

$$\begin{array}{c} I\!R^M \stackrel{F_p}{\to} I\!R^N \\ Input \mapsto Output \end{array} \qquad (1)$$

where the model parameters p come from some parameter domain in $I\!R^P$. Usually there is a large number of input/output observations pairs $(x_j, y_j) \in I\!R^M \times I\!R^N$ such that $y_j = F(x_j)$ holds for the "true" transfer function or model F. The standard identification task is to find a parameter p such that

$$y_j \approx F_p(x_j)$$

holds for all observations $(x_j, y_j) \in I\!R^M \times I\!R^N$, where the parametrized class $\{F_p\}_p$ of models should be adequate for the application in question. Simulation then means the evaluation of $F_p(x)$ for new inputs x. In the context of learning theory the sample of the (x_j, y_j) is called the "training set", and finding p or F_p is called "learning" instead of "identification". From the mathematical point of view one has a nonlinear approximation problem (Braess, 1986), which turns into an interpolation problem, if exact reproduction of the input/output pairs is required. Without losing generality, one can assume the whole process to be scalar–valued, i.e. $N = 1$ in (1). Time series are a special case.

2. KERNELS

In many applications it is useful to map the input observations x_j first into an element $\Phi(x_j)$ of a larger "feature space" describing additional properties of x_j, in order to be able to use a specific distance $dist(\Phi(x_j), \Phi(x_k))$ in feature space that allows to

model the similarity of x_j with x_k much more closely than by the data x_j and x_k themselves. This key idea was termed the "kernel trick" by the community focusing on learning algorithms and "kernel machines" (Schölkopf and Smola, 2002), but its mathematical roots date back to reproducing kernel Hilbert spaces (Aronszajn, 1950; Meschkowski, 1962). The feature map Φ is supposed to map into some function space $\mathcal{F}$ that carries an inner product $(.,.)$ and a reproducing kernel K such that

$$\begin{aligned} \Phi(x) &= K(x,\cdot) \\ (\Phi(x),\Phi(y)) &= (K(x,\cdot),K(y,\cdot)) \\ &= K(x,y) \\ (\Phi(x),v) &= (K(x,\cdot),v) \\ &= v(x) \end{aligned} \tag{2}$$

for all $x, y \in I\!R^M$ and all $v \in \mathcal{F}$. Such kernels exist in many variations (see section 4), and the space $\mathcal{F}$ can be written as the Hilbert space closure of all images $\Phi(x) = K(x,\cdot)$ of the feature map under the above inner product. This "native" Hilbert space intrinsically belongs to the kernel in question, and each Hilbert space of functions with continuous point evaluation has a unique reproducing kernel. Thus there is a one–to–one correspondence between kernels and Hilbert spaces.

3. OPTIMAL MODELS

For modeling purposes there now is a surprising observation: one can find an optimal model without making specific assumptions about the model function F_p and its parametrization. This is what makes kernel techniques interesting for black–box modelling.

Theorem 1. Under all model functions F such that each component of F is an arbitrary function in $\mathcal{F}$ and which reconstructs the observational data exactly, the one with components of minimal norm, if it exists, is necessarily of the form

$$F_\alpha(x) = \sum_j \alpha_j K(x_j, x) \tag{3}$$

with coefficients $\alpha_j \in I\!R^N$.

Our proof sketch of this standard mathematical observation just minimizes the quadratic form (F,F) under the constraints of exact reproduction of the observational data. With Lagrange multipliers α_j one gets the variational equation

$$(F,v) - \sum_j \alpha_j v(x_j) = 0 \text{ for all } v \in \mathcal{F}$$

which leads to (3) via (2). □

In some sense this result eliminates the task of model selection, provided that a feature map and an inner product in feature space is chosen, and if one accepts the resulting minimal norm model.

A drawback of this optimality criterion is that it aims at some kind of "energy" of the model function itself, and not at the prediction quality of the model for new input/output pairs (x,y). But there is another optimality result that precisely aims at reproduction quality.

Theorem 2. Assume that the observational data come from an arbitrary "true" function F from a Hilbert space $\mathcal{F}$ with reproducing kernel K, and consider arbitrary models of the form

$$F_u(x) = \sum_j u_j(x) F(x_j) \tag{4}$$

that are linear in the observational data and use arbitrary weight functions u_j. Then for all x the specific weights $u_j^*(x) \in I\!R$ with

$$K(x,x_k) = \sum_j u_j^*(x) K(x_j,x_k) \tag{5}$$

lead to a minimal worst–case value of the relative error $|F(x) - F_u(x)|/\|F\|$.

The proof sketch for this fact uses Hilbert space duality. Starting from

$$\begin{aligned} |F(x) - F_u(x)|^2 &= |(\delta_x - \sum_j u_j(x)\delta_{x_j})F|^2 \\ &\le \|\delta_x - \sum_j u_j(x)\delta_{x_j}\|^2 \|F\|^2 \end{aligned}$$

and using (2) in the dual form $(\delta_x, \delta_y) = K(x,y)$ one can minimize the above quadratic form with respect to the real variables $u_j(x)$ to get (5) as normal equations

$$(\delta_x, \delta_{x_k}) = \sum_j u_j^*(x)(\delta_{x_j}, \delta_{x_k}). \quad \square$$

The connection of the model (4) with $u_j = u_j^*$ from (5) to the model (3) still needs explanation. If, as discussed in the next section, the matrix with entries $K(x_j,x_k)$ is nonsingular, the functions $u_j^*(x)$ are linear combinations of the $K(x,x_k)$ with the Lagrange interpolation property $u_j^*(x_k) = \delta_{jk}$. Thus the optimal version of (4) is nothing else than a rewriting of (3) in the Lagrange basis, proving that (3) is also minimizing the pointwise reproduction error at all other locations x in the sense of Theorem 2.

The "physical" meaning of the model function (3) can be illustrated by the trivial case

$$F_y(x) = \sum_j y_j \delta_{x_j}(x)$$

for a Dirac delta kernel, and (3) can be seen as a regularized version of the above, still maintaining exact reproduction, but with a different kernel like a Gaussian. Another interpretation views $K(x_j,x)$ as a

similarity measure between the observations x_j and x, being large iff x is close to x_j, and then (3) lets the model return a result that lies close to y_j if x is close to x_j. This line of argument can be made more precise by stochastic assumptions, but one of the goals of this paper is to show that probabilistic arguments are irrelevant to the basics of kernel techniques. They are no more than "add-ons", and this also applies to the nondeterministic parts of learning theory.

The consequence of the arguments of this section is that after proper choice of Φ and K there is no way around working with the representation (3) in black–box modelling. But, after all, picking a very specific Φ and K that suits the application cannot be termed "black–box modelling" any more.

4. POSITIVE DEFINITE KERNELS

If (3) reconstructs the observational data exactly, one has to solve the system

$$F_\alpha(x_k) = y_k = \sum_j \alpha_j K(x_j, x_k) \qquad (6)$$

and this requires that the symmetric matrix with entries $K(x_j, x_k)$ should be nonsingular. Since it is a Gramian due to (2), it must always be positive semidefinite, but one needs additional information to prove its nondegeneracy. Fortunately, the notion of (strictly) positive definite kernels has a long history in mathematics (Stewart, 1976) and provides precisely the required positive definiteness of all such matrices. There also is the notion of conditional positive definiteness (Micchelli, 1986), but this extension is dropped here to simplify the presentation.

By the fundamental Micchelli paper of 1986, there was quite a number of useful (conditionally) positive definite functions on the market, including the radial kernels known as Gaussians, thin–plate splines, polyharmonic splines, and multiquadrics. The term "radial basis function" is used to describe kernels of the form $K(x, y) = \phi(\|x - y\|_2)$ with Euclidean invariance. Special cases are, besides the Gaussian $\exp(-\|x - y\|_2)$, the multiquadrics (Hardy, 1971) and the thin–plate splines (Duchon, 1976).

In 1995 the first compactly supported positive definite radial kernels were constructed (Schaback, 1995*a*; Wu, 1995; Wendland, 1995) and Wendland's functions turned out to be polynomial and of least possible degree for fixed smoothness requirements. Currently, the paper of Schaback and Wendland (2001) surveys the state–of–the–art of construction techniques for positive definite kernels, while other contributions (Schaback, 1999*b*; Schaback, 2000) provide specific properties and relations to integral equations (e.g. "Mercer" kernels in the language of learning theory).

It should be remarked on the side that kernels play a dominant part in integral and differential equations, since they occur in the context of fundamental solutions and Hilbert–Schmidt expansions, for instance. In approximation theory, kernels have a long history dating back to the Dirichlet kernel occurring in Fourier analysis. Finally, it should be kept in mind that one can construct conditionally positive definite kernels K without ever caring for Hilbert space arguments. Applications of kernel models need not care for Hilbert spaces at startup time. But the native Hilbert space for K will always come up later through the back door as a certain closure of the space of functions of the form $K(x, \cdot)$.

5. SUPPORT REDUCTION

If a modeling process involves an abundance of observational data, it is not reasonable to use the model (3) in its original form, because it involves a sum over the full data. Furthermore, in most applications there will be noise in the observations, and then it makes no sense to insist on exact reproduction. However, the two optimality properties of (3) suggest to stay within the overall form of (3), and to try to get away with some useful modifications.

Fortunately, there are strategies to combine both modifications into one. While allowing an error in the reproduction, the complexity of the sum in (3) can be reduced to a smaller set of observational data, called the "support vectors" in the context of learning algorithms. Between the two extreme cases

- using the full data with zero error or
- using no data ($F = 0$) with a huge error

there is a tradeoff between the complexity of a "thinned" sum in (3) and its reproduction quality. This tradeoff is still under investigation, and it remains a major challenge for the future.

The mathematical background for the mainstream of support reduction techniques is based on optimization theory and has nothing to do with modeling, learning theory, time series, probability, or statistics. Reduction simply results from imposing a Chebyshev–type (uniform) bound

$$\|y_j - F(x_j)\|_\infty \le \epsilon \qquad (7)$$

on the reproduction quality, while minimizing a suitable penalty function depending on F, for instance $\|F\|^2$ as in Theorem 1. The necessary Karush–Kuhn–Tucker conditions for the optimum will pick a set of indices j where (7) is attained with equality. This is called the "active set" in optimization theory, and it determines the (hopefully few) "support vectors" and nonzero contributions in (3) using only the indices j from the active set. The other observations are irrelevant for the optimal solution and could have been left out right from the start, if the calculation of the optimal model would ever be repeated.

So far, the modeling problem was stated in "regression" form, but the same argument applies for what is called 'classification" in learning theory. The additional ingredient is the "margin" that plays the part of (7) in a slightly different way that should be explained here for completeness of presentation. The observations x_j are grouped in two classes X^+ and X^-, and the easiest way to define a classification model is to ask for a real–valued function F that does something like

$$\begin{aligned} F(x_j) &\geq +\gamma \quad \text{for all } x_j \in X^+ \\ F(x_j) &\leq -\gamma \quad \text{for all } x_j \in X^- \end{aligned}$$

for some positive γ. Compared with (7) and using standard optimization theory arguments, this is perfectly fine to guarantee a reduction to a few active constraints, if some penalty function on F is minimized. Since the scaling and an additive shift do not matter, one could also ask for an F satisfying

$$\begin{aligned} F(x_j) &\geq 2 \quad \text{for all } x_j \in X^+ \\ F(x_j) &\leq 0 \quad \text{for all } x_j \in X^-. \end{aligned} \tag{8}$$

Unfortunately, the standard approach to classification in learning theory is somewhat more complicated, but essentially a particular case of the above straightforward attack, as is shown now. One looks for a hyperplane in feature space that optimally separates the sets $\Phi(X^+)$ and $\Phi(X^-)$. If a general hyperplane is written as $\{z \ : \ (u,z) = \beta\}$ with an element u of norm 1 in feature space (the normal vector on the hyperplane) and a real number β, then the signed distance of a point z from the hyperplane is $(z,u) - \beta$. Thus one wants a maximal positive "margin" μ with

$$\begin{aligned} (\Phi(x_j),u) - \beta &\geq +\mu \quad \text{for all } x_j \in X^+ \\ (\Phi(x_j),u) - \beta &\leq -\mu \quad \text{for all } x_j \in X^-. \end{aligned}$$

Introducing $z_j := \Phi(x_j)$ and $\sigma_j := \pm 1$ if $x_j \in X^\pm$ this amounts to use optimization to find a maximal positive number μ such that

$$((z_j,u) - \beta)\sigma_j \geq \mu \text{ for all } x_j.$$

The pair (u,β) allows a renormalization, and one can divide the inequalities by μ. Thus the above optimization is equivalent to a minimization of $\|u\|^2$ under the linear Chebyshev–type constraints

$$((z_j,u) - 1)\sigma_j \geq 1 \text{ for all } x_j$$

that allow a reduction argument to active sets based on the Karush–Kuhn–Tucker conditions. The connection to (8) is easily made when taking u as a scalar–valued F and applying (2) in the form

$$(z_j,u) = (z_j,F) = (K(x_j,\cdot),F) = F(x_j)$$

to get

$$(F(x_j) - 1)\sigma_j \geq 1 \text{ for all } x_j,$$

which turns out to be exactly the same as (8). Note that (8) works for any penalty on F, while the the classical geometric margin argument requires a penalty based on $\|F\|$.

6. SPECIAL REDUCTION TECHNIQUES

Besides using Chebyshev–type constraints as in (7) and (8) there are other techniques to reduce the complexity of (3) while allowing some tolerable reproduction error.

"Greedy" methods (DeVore and Temlyakov, 1996; Schaback and Wendland, 2000; Hon *et al.*, 2001) were used to solve (6) partially, working iteratively on the equations where the reproduction error $F_\alpha(x_k) - y_k$ is still too large. It turns out that exact reproduction of a small subset of the observational data often yields small errors on the rest, but research is still incomplete.

The other methods mentioned here are trying to localize the problem somehow. Using fast multipole expansions of kernels (Beatson and Newsam, 1992; Powell, 1993; Beatson and Greengard, 1997; Beatson and Light, 1997; Beatson and Newsam, 1998), one can lump "far" points x_j together and treat them computationally as one, leading to very good reductions in the complexity of solving the system (6) and evaluating the sum in (3). Another localization technique uses partitions of unity (Wendland, 2002) combined with rather arbitrary local models. This technique does not require expansions and allows fairly general applications in modelling. If the partitions of unity satisfy a certain stability property, the global reproduction error can be bounded by the local errors in the subproblems.

It is an interesting open research area to compare and to combine the various reduction techniques. In particular, on can possibly insert greedy, localization, and multipole techniques into advanced methods to solve the huge quadratic linearly constrained problems of section 5 by sophisticated optimization methods.

7. REPRODUCTION QUALITY AND STABILITY

There is a well–established mathematical literature (Duchon, 1978; Madych and Nelson, 1988; Madych and Nelson, 1990; Madych and Nelson, 1992; Wu and Schaback, 1993; Schaback, 1999*a*; Buhmann, 2000) on the error committed by kernel models of the form (3), long before learning machines and black–box modelling with kernels were fashionable. The results will be useful for modelling purposes, but for space limitations only a short summary is possible here, extending an earlier survey (Schaback, 1997).

When experimenting with kernel models (3) and corresponding systems (6) it turns out that results often depend crucially on the scaling of the kernel in relation to the density of the observational data. To quantify the latter, the fill distance

$$h_{X,\Omega} := \sup_{y \in \Omega} \min_{x_j \in X} \|y - x_j\|_2$$

of the set $X = \{x_j\}$ of observational data within an enclosing domain Ω is useful. Then one looks

at sequences of observational data such that the fill distance $h_{X,\Omega}$ tends to zero, and the goal is to prove convergence in the sense that the reproduction error tends to zero as a function of $h_{X,\Omega}$. If the kernel K is fixed throughout this process, the situation is called "nonstationary" within approximation theory, while a "stationary" setting scales the kernel with $h_{X,\Omega}$ to keep the data distance and the kernel width proportional.

For the stationary setting, theory (Buhmann, 1989) says that integrable kernels like the Gaussian, the inverse multiquadric or the compactly supported Wendland kernels cannot yield convergence, though in practice one often observes that the errors are small enough to keep the user satisfied. In fact, the error usually decreases if the ratio of the kernel width and the fill density (i.e. the "bandwidth" in the matrix of (6)) is increased, and this is sufficient in many cases to keep the error below a tolerable level.

In the nonstationary setting the reproduction error always behaves like a power $h_{X,\Omega}^k$ where $k > 0$ increases with the smoothness of K. If the kernel is analytic (this occurs for the Gaussian and for multiquadrics, for instance), the error decreases even exponentially like $\exp(-c/h_{X,\Omega})$ with a positive constant c, at least in theory (Madych and Nelson, 1992).

However, this fantastic convergence behavior comes at a price. In practice, the linear systems (6) get more and more ill–conditioned (Narcowich and Ward, 1991; Narcowich and Ward, 1992) when $h_{X,\Omega}$ gets small, and this effect is dramatically increasing with the smoothness of K. It can be proven (Schaback, 1995*b*) that good reproduction always comes with instability and vice versa, while additional smoothness of the kernel boosts both of them, unfortunately. If the user does not apply additional techniques like preconditioning or localization, the best choice of scale usually is the one that works close to the condition limits of the machine. Thus preconditioning is another important topic (Dyn *et al.*, 1983; Jetter and Stöckler, 1995; Beatson *et al.*, 1999; Fasshauer and Jerome, 1999; Hon and Kansa, 2000) in the context of making (3) work for large–scale application problems. Even for Gaussians, which show extremely good reconstruction quality and catastrophic instability, there was no well–established preconditioning technique so far, but the situation will improve (Schaback, 2002).

The above discussion was restricted to exact reconstruction of data from functions in the Hilbert space associated to the kernel. This "native" Hilbert space is very small when the kernel is very smooth, and thus in theory the user must make sure to use a kernel that is not too smooth. However, if the reconstruction is allowed to be not exact, or if there is contamination by noise, one can observe in practice and prove (Schaback, 1996) that the convergence rate in the nonstationary setting adapts in an optimal and local way to the smoothness of the unknown function supplying the observational data.

This is a partial result concerning the tradeoff between complexity of (3) and the reproduction quality, but the proof technique, when studied in detail, only treats the case where the actually used observational data still "covers" the whole data domain Ω. Compared to the reduction technique of section 5 this is a worst–case scenario that does not exploit specific features of the observational data.

8. ACKNOWLEDGEMENT

Special thanks go to Alexander Hornstein and Ulrich Parlitz for help with a first version of this manuscript.

9. REFERENCES

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337–404.

Beatson, R.K. and G.N. Newsam (1992). Fast evaluation of radial basis functions: I. Advances in the theory and applications of radial basis functions. *Comput. Math. Appl.* **24**(12), 7–19.

Beatson, R.K. and G.N. Newsam (1998). Fast evaluation of radial basis functions: moment based methods. *SIAM J. Comput.* **19**, 1428–1449.

Beatson, R.K. and L. Greengard (1997). A short course on fast multipole methods. In: *Wavelets, Multilevel Methods and Elliptic PDEs* (M. Ainsworth, J. Levesley, W. Light and M. Marletta, Eds.). pp. 1–37. Oxford University Press.

Beatson, R.K. and W.A. Light (1997). Fast evaluation of radial basis functions: Methods for two–dimensional polyharmonic splines. *IMA Journal of Numerical Analysis* **17**, 343–372.

Beatson, R.K., J.B. Cherrie and C.T. Mouat (1999). Fast fitting of radial basis functions: Methods based on preconditioned GMRES iteration. *Advances in Computational Mathematics* **11**, 253–270.

Braess, D. (1986). *Nonlinear Approximation Theory*. Springer, Berlin.

Buhmann, M.D. (1989). Multivariable interpolation using radial basis functions. PhD thesis. University of Cambridge.

Buhmann, M.D. (2000). Radial basis functions. *Acta Numerica* **10**, 1–38.

DeVore, R. A. and V. N. Temlyakov (1996). Some remarks on greedy algorithms. *Advances in Computational Mathematics* **5**, 173–187.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *Rev. Française Automat. Informat. Rech. Opér. Anal. Numer.* **10**, 5–12.

Duchon, J. (1978). Sur l'erreur d'interpolation des fonctions de plusieurs variables pas les D^m–splines. *Rev. Française Automat. Informat. Rech. Opér. Anal. Numer.* **12**(4), 325–334.

Dyn, N., D. Levin and S. Rippa (1983). Surface interpolation and smoothing by "thin plate" splines. *Journal of Approximation Theory* **38**, 445–449.

Fasshauer, G. and J. W. Jerome (1999). Multistep approximation algorithms: Improved convergence rates through postconditioning with smoothing kernels. *Advances in Computational Mathematics* **10**(1), 1–27.

Hardy, R.L. (1971). Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **76**, 1905–1915.

Hon, Y.C. and E.J. Kansa (2000). Circumventing the ill-conditioning problem with multiquadric radial basis functions: applications to elliptic partial differential equations. *Comput. Math. Applic.* **39**, 123–127.

Hon, Y.C., R. Schaback and X. Zhou (2001). Adaptive greedy algorithms for solving large RBF collocation problems. manuscript.

Jetter, K. and J. Stöckler (1995). A generalization of de Boor's stability result and symmetric preconditioning. *Advances in Computational Mathematics* **3**, 353–367.

Madych, W.R. and S.A. Nelson (1988). Multivariate interpolation and conditionally positive definite functions. *Approximation Theory and its Applications* **4**, 77–89.

Madych, W.R. and S.A. Nelson (1990). Multivariate interpolation and conditionally positive definite functions II. *Mathematics of Computation* **54**, 211–230.

Madych, W.R. and S.A. Nelson (1992). Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. *Journal of Approximation Theory* **70**, 94–114.

Meschkowski, H. (1962). *Hilbertsche Räume mit Kernfunktion*. Springer, Berlin.

Micchelli, C.A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation* **2**, 11–22.

Narcowich, F.J. and J.D. Ward (1991). Norm of inverses and condition numbers for matrices associated with scattered data. *Journal of Approximation Theory* **64**, 69–94.

Narcowich, F.J. and J.D. Ward (1992). Norm estimates for the inverses of a general class of scattered–data radial–function interpolation matrices. *Journal of Approximation Theory* **69**, 84–109.

Powell, M.J.D. (1993). Truncated Laurent expansions for the fast evaluation of thin plate splines. *Numer. Algorithms* **5**(1–4), 99–120.

Schaback, R. (1995*a*). Creating surfaces from scattered data using radial basis functions. In: *Mathematical Methods for Curves and Surfaces* (T. Lyche M. Daehlen and L.L. Schumaker, Eds.). pp. 477–496. Vanderbilt University Press, Nashville, TN.

Schaback, R. (1995*b*). Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics* **3**, 251–264.

Schaback, R. (1996). Approximation by radial basis functions with finitely many centers. *Constructive Approximation* **12**, 331–340.

Schaback, R. (1997). On the efficiency of interpolation by radial basis functions. In: *Surface Fitting and Multiresolution Methods* (A. LeMéhauté, C. Rabut and L.L. Schumaker, Eds.). pp. 309–318. Vanderbilt University Press, Nashville, TN.

Schaback, R. (1999*a*). Improved error bounds for scattered data interpolation by radial basis functions. *Mathematics of Computation* **68**, 201–216.

Schaback, R. (1999*b*). Native Hilbert spaces for radial basis functions I. In: *New Developments in Approximation Theory* (M.D. Buhmann, D. H. Mache, M. Felten and M.W. Müller, Eds.). pp. 255–282. Number 132 In: *International Series of Numerical Mathematics*. Birkhäuser Verlag.

Schaback, R. (2000). A unified theory of radial basis functions (native Hilbert spaces for radial basis functions II). *J. Comp. Appl. Math.* **121**, 165–177.

Schaback, R. (2002). Multivariate interpolation by polynomials and radial basis functions. Manuscript.

Schaback, R. and H. Wendland (2000). Adaptive greedy techniques for approximate solution of large RBF systems. Preprint.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels*. MIT Press.

Sjoberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Glorennec, H. Hjalmarsson and A. Juditsky (1995). Nonlinear black-box modeling in system identification: a unified overview.

Stewart, J. (1976). Positive definite functions and generalizations, an historical survey. *Rocky Mountain J. Math.* **6**, 409–434.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics* **4**, 389–396.

Wendland, H. (2002). Fast evaluation of radial basis functions: Methods based on partition of unity. In: *Approximation Theory X: Wavelets, Splines, and Applications* (C. K. Chui, L. L. Schumaker and J. Stöckler, Eds.). pp. 473–483. Vanderbilt University Press.

Wu, Z. (1995). Multivariate compactly supported positive definite radial functions. *Advances in Computational Mathematics* **4**, 283–292.

Wu, Z. and R. Schaback (1993). Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis* **13**, 13–27.