Advances in Computer Methods for Partial Differential Equations VII
R. Vichnevetsky, D. Knight and G. Richter (Editors)

SELECTION OF A RADIAL BASIS FUNCTION FOR DATA INTERPOLATION

Donald E. Myers
Department of Mathematics
University of Arizona
Tucson, AZ 65721
myers@math.arizona.edu
(602) 621-6859

ABSTRACT

The radial basis function interpolator belongs to a general class of interpolators which include the thin plate and smoothing spline. The latter two are derivable by imposing a smoothness condition on the interpolating function and correspond to a specific choice of the kernel or structure function but these choices are not unique. It is well-known that positive definiteness of an appropriate form is sufficient to ensure that the system of equations, determining the coefficients in the radial basis function representation, will have a unique solution. Hence any valid choice of the kernel will produce an exact interpolator, i.e., the interpolated surface will pass through the data points. The "optimal" choice of the kernel function is not determined by a least-squares fit of the interpolating surface to the data and an alternative measure of the fit is needed.Imposing such a condition may be justified by a priori knowledge of the function to be interpolated. By using an alternative but equivalent formulation of the interpolator the exactness property is used to evaluate the fit of the kernel function to the data. There are several natural measures of the collective closeness of the pairs. These include the mean error, the normalized mean square error (normalized by the minimized variance of the interpolation error), the correlation of the data vs the interpolated values, the correlation of the interpolated values vs the normalized interpolation errors, and the frequency distribution of the normalized errors. These are usually called cross-validation statistics. The applicability of cross-validation to the choice of the kernel for radial basis function interpolation is demonstrated.

INTRODUCTION

Let $x_1,....x_n$ be in R^q and $f_0(x),...f_p(x)$ real valued linearly independent functions and $g(x,a)$ a symmetric function defined on R^q x R^q. Consider an interpolator of an unknown function y of the form

$$y^*(x) = b_1g(x,x_1) + ... + b_ng(x,x_n) \qquad (a)$$

There are two general conditions that can be imposed on y^*, one is exactness, i.e., the interpolated values at data points match the data and secondly, that y^* is "close" to the data values at the data points. The latter condition is usually imposed when another condition such as maximal smoothness is required although that is not the unique choice. Exactness leads to the system of equations

$$b_1g(x_j,x_1) + ... + b_ng(x_j,x_n) \qquad ; j=1,...,n \qquad (b)$$

where $y(x_1),....y(x_n)$ are the values of the unknown function y. Although the coefficient matrix in (b) is square it may not be invertible. It is known that for some choices of g that the invertibility problem is solved if (a) is replaced by

$$y^*(x) = b_1g(x,x_1) + ... + b_ng(x,x_n) + a_0f_0(x)+ ... +a_pf_p(x) \qquad (1)$$

Invoking the Exactness condition gives the system

$$b_1g(x_j,x_1) + ... + b_ng(x_j,x_n) + a_0f_0(x_j)+ ... +a_pf_p(x_j) = y(x_j) \qquad ; j=1,...,n \qquad (2)$$

However in that case the coefficient matrix is not square and additional conditions of the form

$$b_1 f_k(x_1) + \ldots + b_n f_k(x_n) = 0;\ k = 0,\ldots,p \qquad (3)$$

There are two special cases that are of interest, first the f's might be polynomials in the coordinates of x and secondly g might be function of x-y or in particular a function of the magnitude of x-y. As noted by Micchelli [1], even with these modifications this system may not have a unique solution. Micchelli has shown that conditional positive definiteness of g with respect to the polynomial functions is a sufficient condition for invertibility. Micchelli's results will extend to the more general class of f's provided that g is conditionally positive definite with respect to the set of f's, this result is given in Myers [2,3]. If the system (2) is modified slightly as in Cressie [4] then a smoothing interpolator is obtained. The thin plate and smoothing splines are special cases. As shown in Myers[2], [3] y^* can be re-written in an equivalent form

$$y^*(x) = c_1(x)y(x_1) + \ldots + c_n(x)y(x_n) \qquad (1')$$

and the system (2) is replaced by

$$c_1(x)g(x_j,x_1) + \ldots + c_n(x)g(x_j,x_n) + d_0(x)f_0(x_j) + \ldots + d_p(x)f_p(x_j) = g(x,x_j) \qquad ;\ j=1,\ldots,n \qquad (2')$$

and (3)

$$c_1(x)f_k(x_1) + \ldots + c_n(x)f_k(x_n) = f_k(x);\ k=0,\ldots,p \qquad (3')$$

The alternative form is obtained by considering y as a realization of a random function Y and requiring that y^* be the minimum variance, unbiased LINEAR predictor of Y(x). This form of the interpolator has been derived by a number of authors including Goldberger [5] and Matheron [6] although Matheron is considerably more general. Matheron [7,8] has shown the equivalence of the form given by (1'), (2') and (3') with the thin plate spline. Watson [9] gave a more rudimentary derivation of this equivalence and in the case of the thin plate spline shows the relationship between the kernel and the Green's function of a certain differential operator. In the case that q=1, Dolph and Woodbury [10] have given a very general result linking the covariance of a second order stationary process and the Green's Function of a of a particular form of differential operator. When g is the covariance of a second order stationary random function there there is a distance (called the range) such that for two points at a greater distance the covariance is constant (or the covariance is asymptotically constant in which case there is an effective range. Given a finite number of data locations there exists a neighborhood sufficiently large such that for any location x outside the neighborhood, the value of $g(x, x_i)$ is a constant (or nearly so). If further at least one of the f's is a constant function then the sum of the b's in (1) is zero. It is seen then that when (1) is used as an EXtrapolator, its behavior is largely determined by the f's. Moreover the sum in (1) involving the f's is the unbiased estimator of the mean of Y at x. When p=0 and $f_0 = 1$ then the mean is a constant and is estimated by a weighted sum of the data values. If in addition the data locations are sufficiently spread out then the weighted sum is just the arithmetic sum. Some characteristics of the interpolator are more easily seen or described by using (1'). As the distance from x to a data location x_i increases, the weight $c_i(x)$ diminishes. That is, the location x_i has less "influence" on the interpolation at x. In practice then it is common to use only the nearest data locations for interpolation at a particular point. In (1) the coefficients do not depend on x but in (1') they do and hence (1') is somewhat easier to use. Since any valid g, i.e., any g having the requisite positive definiteness property will produce an exact interpolator, it is necessary to impose additional conditions to obtain a unique or optimal choice of g. In the case of the thin plate spline g is uniquely determined by imposing the maximal smoothness condition. In some instances this may be too strong a condition or in the case when y is unknown the appropriate degree of smoothness may be indeterminate. By combining the exactness property and the use of moving neighborhoods it is possible to quantify the fit of a particular kernel to the data.

COMPARISON OF THE TWO FORMS

When the interpolator is written in the radial basis function form given by (1) the coefficients depend only on the kernel, the sample location pattern and the data values. The differentiability properties including smoothness are all captured in the kernel function as it appears in the interpolator (1). In this form it is somewhat less obvious that a moving neighborhood might be used but it has the advantage that it provides a clear functional representation for

the interpolator which would be convenient for plotting. While the equations given in (2) are seen to be the exactness condition it is not as obvious as to the purpose of the conditions in (3) except that they are sufficient for the system to have a unique solution. This form of the interpolator does clearly show the decomposition of y^* into the "interpolator" and the "extrapolator".

In contrast the form given by (1') clearly shows the dependence on the data values but the dependence on x is hidden in the coefficients. The differentiability and smoothness properties of y^* are somewhat hidden in the coefficients. As x changes a new set of coefficients must be computed. In practice one identifies a set of locations where f is to be estimated or interpolated. Unless the sample location pattern in the moving neighborhood remains fixed as x changes it is necessary to re-solve the system for each choice of x. While the problem is simplified somewhat by using a unique neighborhood, i.e., all data locations are used to estimate at all points and hence the coefficient matrix must be inverted only once, this is in general not practical when the data set is large. In the geostatistical literature it is common practice to use only the nearest locations, usually less than 25. Since in (1') the weights on the data values are directly utilized it is easy to classify the important data locations when estimating or interpolating at a particular point x. Perhaps the most important reason for utilizing the alternative form is that it is easily derived from the random function model and the conditions given in (3') correspond to the unbiasedness. As a by-product it is natural to consider jackknifing the data locations and to consider interpolation at a data location when that data location is suppressed. This is the basis for what is called cross-validation and it provides a means for evaluating the fit of the kernel to the data.

The interpolation problem can be generalized in another way. Rather than estimating/interpolating the value of y at a point, estimate the value of a linear functional. The value at the point x is simply a special case. Another important case is given by spatial integrals. Let V be a volume then consider the average value of y over V. The alternative form (1') will not change, only the right sides of the equations in (2') will change being replaced by average values. If the form given by (1) is used then the estimator itself must change.

There is one additional advantage of the alternative form that is worth mentioning. Both forms are essentially independent of the dimension of the domain space R^k but it is much simpler to extend the alternative form to the case where Y has values in R^m. In particular there is a natural way to quantify the interrelationship between the components of the vector valued Y by the use of cross-covariances or cross-variograms. This extension is discussed in Myers [3,11].

In the remainder of the paper we assume that g(x,u) is a function of x-u and hence we write simply g(x-u).

ERROR CHARACTERIZATION

The system of equations (2'), (3') is obtained by requiring that (1') be an unbiased, minimum variance interpolator, that is, the coefficients in (1') are chosen so that $\mathrm{Var}\{Y^*(x)-Y(x)\}$ is minimized. Using the resulting system of equations the minimized variance (called the kriging variance in the geostatistical literature) is given by

$$c_1(x)g(x,x_1) + \ldots + c_n(x)g(x,x_n) - d_0(x)f_0(x) + \ldots + d_p(x)f_p(x) \qquad (4)$$

DIRECT ESTIMATION OF THE KERNEL

One of the advantages of the alternative formulation given by (1'), (2') and (3') is that the kernel is a generalized covariance as defined in Matheron [5]. To illustrate the method we consider only the cases where g is a covariance or a variogram. A covariance is positive definite whereas if g is a variogram then -g is conditionally positive definite of order zero. In addition if g is a covariance then g(0)-g(x) is a variogram hence it is sufficient to consider estimation of variograms. More explicitly the random function F is assumed to satisfy the Intrinsic Hypothesis

(i) $E\{Y(x+u)-Y(x)\} = 0$ for all x, u

(ii) $0.5\mathrm{Var}\{Y(x+u)-F(x)\} = g(x)$ exists and depends only on x

That is, the first order increments of Y are second order stationary. Note that it is not necessary to assume that g is isotropic (a function of the magnitude of x only). It is common to model geometric anisotropies in variograms. From (i) and (ii) it is easy to see that

$$g^*(x) = (1/N)\sum [y(x_i)-y(x_j)]^2 \qquad (5)$$

is an unbiased estimator of g(x), the sum is taken over all pairs of data locations such that x_i-x_j =x. In practice, as is described in Myers[12,13], it is necessary to consider distance classes and directional windows especially in the case of irregular data grids. Unfortunately this estimator only produces estimates of g for certain values of x and it is necessary to know g in functional form. The practical solution is to consider positive linear combinations of known valid kernels. A plot of g^* as a function of the magnitude of x and for a particular direction will indicate certain important parameters as well as model types. There may be a discontinuity at the origin, called the "nugget effect" and if the plot levels off and becomes approximately constant then the height is the sill and the distance at which this occurs is the range. In variogram form the nugget effect model is given by

$$g(0) = 0 \quad \text{and} \quad g(x) = \sigma^2 \quad \text{for } x = 0.$$

In the case of Gaussian or Exponential models there is an effective range. If the data is noisy and an exact interpolator is used then the nugget effect will include the variance of the noise term, more generally it represents a spatial correlation structure at a distance less than the minimum intersample location distances. There are several advantages to using a variogram estimator in lieu of a covariance estimator. A variogram can be unbounded and this will not be detected with a covariance estimator. Estimation of the covariance requires separate estimation of the mean. In general the sample variogram given by g^* will be somewhat noisy and is known to be an imperfect estimator of the the variogram. There are various reasons for this. A regular grid does not produce a sufficient number of pairs for short lag distances unless the mesh distance is quite small whereas it is the values of g for short distances that are the most important. There are empirical results for the use of least squares fitting to valid models and for the use of maximum likelihood but the latter in particular is dependent on a strong multivariate normality assumption for Y. In practice the fitting is done visually and requires both experience and a knowledge of the function or phenomenon being interpolated. This visual fitting is frequently combined with the cross-validation technique discussed later.

There is an additional complication that should be mentioned. In the case where condition (i) is not satisfied but rather a weaker condition

$$(i') \quad E[Y(x)] = t_0 f_0(x) + \ldots + t_p f_p(x)$$

where the t_j's are unknown coefficients then g^* is not an unbiased estimator of g. Unfortunately this leads to a circular problem without a fully satisfactory solution. While Y(x) might not satisfy (i), Y(x)- E[Y(x)] would and hence it is necessary to estimate the unknown coefficients. The optimal estimation of those coefficients requires knowing g which can not be estimated or modeled until the coefficients are estimated. This problem has been extensively considered in the geostatistical literature.

SMOOTHNESS DETERMINANTS

As noted above the differentiability properties of f^* are easily seen to be those of the kernel g when (1), (2. and (3) are used, *provided* that all data locations are used at one time so that the coefficients in (1) are computed only once. If an additional term $b_i\sigma^{-1}$ is added to the left side of (2) or $c_j(x)\sigma^{-1}$ is added to the left side of (2') then the interpolator is no longer exact. In the context of (1'), (2') and (3') this corresponds to assuming that the random function Y is replaced by Z(x) = Y(x) + e(x) where e(x) is a noise term. The noise term is assumed to be uncorrelated with respect to Y(x) and to have a pure nugget variogram. Then the data values $y(x_1),\ldots,y(x_n)$ in (2) and (1') are replaced by $z(x_1),\ldots,z(x_n)$. y^* is then a smoothing interpolator and the parameter σ^{-2} which could be interpreted as the variance of the noise term represents the degree of smoothing relative to the smoothness characteristics of the kernel g. When a pure nugget effect variogram is used the interpolator given by (1') differs from a polynomial trend surface only at the data locations.

CROSS-VALIDATION

Sequentially, one at a time, the data locations are suppressed and an interpolated value is obtained using only the remaining data locations (or only those in a neighborhood about the suppressed location). At each data location x_i there will be three pieces of information, the data value $y(x_i)$, the interpolated value $y^*(x_i)$ and the minimized estimation variance σ_i^2. Note that the latter two may be somewhat sensitive to the choice of the moving neighborhood and the limitation on the number of data locations used in the interpolation. For a "well-chosen" kernel, i.e., variogram,

the interpolated values should be close to the corresponding data values where close is in the sense of the minimized estimation variance. While the general question of choosing the kernel is not well-posed (given only a finite number of data locations), choices of the kernel and/or of the parameters of the kernel can be ranked with respect to the closeness of the interpolated values to the data values. There are at least five statistics for quantifying this closeness, as shown in Myers [11]. These include the following:

$$E_1 = (1/n)\{(y(x_1)-y^*_1(x_1)) + \ldots + (y(x_n)-y^*_n(x_n))\}$$

$$E_2 = (1/n)\{(y(x_1)-y^*_1(x_1))^2 + \ldots + (y(x_n)-y^*_n(x_n))^2\}$$

$$E_3 = (1/n)\{(y(x_1)-y^*_1(x_1))^2/\sigma^2_1 + \ldots + (y(x_n)-y^*_n(x_n))^2/\sigma^2_1\}$$

E_4 = sample correlation $\{(y(x), y^*(x))\}$

E_5 = sample correlation $\{(y(x)-y^*(x)),y^*(x))\}$

In addition it common practice to construct the histograms of the errors and of the normalized errrors. Under moderate assumptions it is possible to compute the expected values of these statistics. The expected value of E_1 should be zero and hence one would expect the sample value to be close to zero. The expected value of E_2 is the average of the minimized estimation variances while the mean of E_3 should be one. If Y is assumed to second order stationary then the variance of E_1 can be computed but the variances of all the rest either require strong distributional assumptions (such as multivariate normality) or knowledge of fourth order moments of Y. Under the second order stationarity assumption the expected values of E_4, E_5 respectively should be close to one and zero. They are in general only close because of the presence of the Lagrange multipliers in (2*). In addition to computing the sample correlations it is useful to construct scatter plots. In some instances it is preferable to use the normalized error in E_3 in lieu of the estimation error. Note that a form of cross-validation is used in connection with the smoothing parameter in a smoothing spline, that is rather different from the form described above. The smoothing spline is not exact and the kernel is assumed already determined hence in that case the cross-validation is not utilized for kernel selection or ranking.

THE USE OF CROSS-VALIDATION

While cross-validation provides some measure of the goodness-of-fit of a particular kernel, it is most useful in comparing different choices or different parameter values. There are a number of known variogram and covariance models and in general there are two parameters associated with a model. It is easiest to describe the isotropic models first and then describe how to incorporate geometric anisotropies into those models. Some of the standard models are described in greater detail in Myers [11,12]. If Y is second order stationary then the value of the covariance at distance zero is called the SILL, alternatively it is the constant value that the variogram achieves or approaches asymptotically. The distance at which the covariance is zero is called the RANGE, in the case of the Exponential and Gaussian models there is only an effective range. Any positive linear combination of valid models is again a valid model, meaning that the appropriate positive definiteness condition is satisfied. Some variograms do not correspond to covariances and hence do not have sills or ranges but there will be corresponding parameters. Cross-validation can then be used to evaluate such linear combinations by varying the parameter choices.

In examining the interpolator given by (1) it is easy to see that for a fixed point x, the values of $y^*(x)$ depend on 1. the kernel function (and its parameters), 2. the data values, 3. the data location pattern, 4. the relationship of the point x to the data location pattern. If a moving neighborhood is used then both 3. and 4. change as x changes. When applying cross-validation however one can fix 2., 3. and 4. hence only the variogram (or its parameters) changes although one can consider different search neighborhoods. The continuity of y^* relative to the variogram is sensitive to the search neighborhood and in turn this is sensitive to the smoothness of the unknown function. Cross validation then will be used both to rank the variogram choices and also to identify unusual features of the data set. These two objectives are not totally separable.

SUMMARY

A radial basis function interpolator is generated by any kernel with the appropriate positive definiteness property and hence is not uniquely determined. When the principal information concerning the function to be interpolated consists of the data values, it is important to fit the kernel, e.g., the interpolator to the data. By using an alternative but equivalent form of the interpolator and the exactness property, cross-validation statistics allow ranking choices of the kernel with respect to their fit to the data.

REFERENCES

1. C. Micchelli, "Interpolation of Scattered Data:Distance Matrices and Conditionally Positive Definite Functions" Constructive Approximation Vol 2, pp. 11-22, 1986

2. D.E. Myers, "Interpolation with Positive Definite Functions" Sciences de la Terre Vol 28, pp. 252-265, 1988

3. D.E. Myers, "Interpolation, Estimation Methods and Positive Definiteness" to appear in An International Journal of Computers and Mathematics with Applications, 1992

4. N. Cressie, "Reply to a Letter by G. Wahba" The American Statistician Vol 44, pp.256-258, 1990

5. A.S. Goldberger, "Best Linear Unbiased Prediction in the Generalized Regression Model" J. American Stat. Assn pp.369-375, 1966

6. G. Matheron, "The Intrinsic Random Functions and their Applications" Adv. Applied Prob. Vol. 5, pp.439-468, 1973

7. G. Matheron, Splines et Krigeage:Leur Equivalence Formelle Internal Report N-667, Centre de Geostatistique, Fontainebleau, 26p

8. G. Matheron, Remarques sur le Krigeage et son dual Internal Report N-695, Centre de Geostatistique, Fontainebleau, 36p

9. G. Watson, "Smoothing and Interpolation and Splines" Math. Geology Vol 16, pp.601-615, 1984

10. C.L. Dolph and M.A. Woodbury, "On the Relation between Green's Functions and Covariances of Certain Stochastic Processes and its Application to Unbiased Linear Prediction" Trans. Amer. Math. Soc. Vol 72, pp.519-550, 1952

11. D.E. Myers, "Multivariate, Multidimensional Smoothing" in Spatial Statistics and Imaging, Proc. of the AMS-IMS-SIAM Joint Summer Research Conference on Spatial Statistics, June 18-24, 1988, Bowden College, Lecture Notes-Monograph Series, Inst. Math. Statistics, Hayward, pp. 275-285, 1991

12. D.E. Myers, "On Variogram Estimation" in Proceedings of the First Inter. Conf. Stat. Comp., Cesme, Turkey, 30 March-2 April, 1987, American Sciences Press, Vol II, pp.261-281, 1991

13. D.E. Myers, "Interpolation of Spatially Located Data" Chemometrics and Intelligent Laboratory Systems, Vol 11, pp.209-228, 1991