Radial Basis Functions: nagdmc_rbf

Purpose

nagdmc_rbf fits a radial basis function (RBF) model to data records with p independent variables.

Declaration

```
#include <nagdmc.h>
void nagdmc_rbf(long rec1, long nvar, long nrec, long dblk, double data[],
                long nxvar, long xvar[], long yvar, long nrbf, double cen[],
                int dtype, int mtype, double m[], int rbftype, double alpha,
                double tol, double lambda[], double w[], double yhat[],
                double model[], int *info);
```

Parameters

1: **rec1** – `long` *Input*

On entry: the index in the data of the first data record used in the analysis.

Constraint: $\textbf{rec1} \geq 0$.

2: **nvar** – `long` *Input*

On entry: the number of variables in the data.

Constraint: $\textbf{nvar} > 1$.

3: **nrec** – `long` *Input*

On entry: the number of consecutive records, beginning at **rec1**, used in the analysis.

Constraint: $\textbf{nrec} > 1$.

4: **dblk** – `long` *Input*

On entry: the total number of records in the data block.

Constraint: $\textbf{dblk} \geq \textbf{rec1} + \textbf{nrec}$.

5: $\textbf{data}[\textbf{dblk} * \textbf{nvar}]$ – `double` *Input*

On entry: the data values for the jth variable (for $j = 0, 1, \ldots, \textbf{nvar}-1$) are stored in $\textbf{data}[i*\textbf{nvar}+j]$, for $i = 0, 1, \ldots, \textbf{dblk} - 1$.

6: **nxvar** – `long` *Input*

On entry: the number of independent variables. If $\textbf{nxvar} = 0$ then all variables in the data, excluding **yvar**, are treated as independent variables.

Constraint: $0 \leq \textbf{nxvar} < \textbf{nvar}$.

7: $\textbf{xvar}[\textbf{nxvar}]$ – `long` *Input*

On entry: the indices indicating the position in **data** in which values of the independent variables are stored. If $\textbf{nxvar} = 0$ then **xvar** must be 0, and the indices of independent variables are given by $j = 0, 1, \ldots, \textbf{nvar} - 1$; $j \neq \textbf{yvar}$.

Constraints: if $\textbf{nxvar} > 0$, $0 \leq \textbf{xvar}[i] < \textbf{nvar}$, for $i = 0, 1, \ldots, \textbf{nxvar} - 1$; otherwise **xvar** must be 0.

8: **yvar** – `long` *Input*

On entry: the index in **data** in which values of the dependent variable are stored.

Constraints: $0 \leq \textbf{yvar} < \textbf{nvar}$; if $\textbf{nxvar} > 0$, $\textbf{yvar} \neq \textbf{xvar}[i]$, for $i = 0, 1, \ldots, \textbf{nxvar} - 1$.

9: **nrbf** – `long` *Input*

On entry: the number of RBFs in the model.

Constraint: $\textbf{nrbf} \geq 1$.

10: **cen**[**nrbf**$*p$] – `double` *Input*

On entry: the locations in the input space of **nrbf** RBFs, stored by row. The centre location of the kth RBF on the jth independent variable is given by **cen**$[k*p+j]$, for $j=0,1,\ldots,p-1$; for $k=0,1,\ldots,$**nrbf**-1.

11: **dtype** – `int` *Input*

On entry: the value of **dtype** describes the distance function used. If **dtype** $=0$, the ℓ_2-norm or Euclidean distance is used; otherwise **dtype** $=1$, and the ℓ_2-norm or Manhattan distance is used.

Constraint: **dtype** $\in\{0,1\}$.

12: **mtype** – `int` *Input*

On entry: the value of **mtype** determines the scaling type used to compute distances; valid options are:

0: user-supplied scalar;
1: scale using standard deviations of data;
2: user-supplied scalings.

The Euclidean distance has the additional option:

3: Mahalanobis distances.

Constraint: if **dtype** $=0$, **mtype** $\in\{0,1,2,3\}$; otherwise **mtype** $\in\{0,1,2\}$.

13: **m**$[d]$ – `double` *Input*

On entry: an array of d user-supplied scalings used to compute the radial distances. The value of d depends on the value of **mtype**. If **mtype** $=0$, $d=1$; if **mtype** $=2$, $d=p$; otherwise **m** must be 0.

Constraint: if **mtype** $\in\{0,2\}$, **m**$[j]>0.0$ contains the jth user-supplied scaling value, for $j=0,1,\ldots,d-1$.

14: **rbftype** – `int` *Input*

On entry: the value of **rbftype** determines the kind of radial basis function used in the model:

0: linear
1: cubic
2: thin plate spline
3: Gaussian
4: multiquadric
5: inverse multiquadric
6: Cauchy

Constraint: **rbftype** $\in\{0,1,2,3,4,5,6\}$.

15: **alpha** – `double` *Input*

On entry: if **rbftype** $\in\{4,5,6\}$, the value of the RBF parameter α; otherwise **alpha** is not referenced.

Constraint: if referenced, **alpha** >0.0.

Suggested value: if referenced, **alpha** $=1.0$.

16: **tol** – `double` *Input*

On entry: the value of **tol** used to determine the tolerance for setting eigenvalues equal to zero in the singular value decomposition. If **tol** is less than the machine accuracy ϵ, **tol** will be set equal to ϵ.
Suggested value: **tol** $=1\times10^{-6}$.

17: **lambda** – `double` *Input*

On entry: the value of the the ridge regression parameter, λ, in the penalised sum of squares error functon. Setting: **lambda** $=0$ gives the ordinary least squares solution.

Constraint: **lambda** ≥ 0.0.

18: **w**[1+**nrbf**] – `double` *Output*

On exit: **w**$[k]$ contains the value of the scalar multiplier on the kth RBF, for $k=0,1,\ldots,$**nrbf**-1; the intercept value is stored in **w**[**nrbf**].

19: **yhat[nrec]** – `double` *Output*

On exit: **yhat**$[i]$ contains the RBF model approximation for the dependent variable in the ith data record, for $i = 0, 1, \ldots,$ **nrec** $- 1$.

20: **model**$[6 + (p + 1)($**nrbf**$+1) + d]$ – `double` *Output*

On exit: contains information that specifies the RBF model for use in **nagdmc_predict_rbf**, where $d = 1$ if **mtype** $= 0$; $d = p * (p + 1)/2$ if **mtype** $= 3$; and otherwise $d = p$. If **model** is 0, it is not referenced and model summary information is not returned.

21: **info** – `int *` *Output*

On exit: **info** gives information on the success of the function call:

i; $i = 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, \ldots, 15, 17$: the specification of the ith formal parameter was incorrect.

41: computation of the pseudo inverse failed; increasing the value of λ may avoid this error return.

42: an error occurred computing the scalings.

99: the function failed to allocate enough memory.

100: an internal error occurred during the execution of the function.

Notation

nrec	the number of data records in the analysis, n.
data	the data values, X.
nxvar	determines the number of independent variables in the analysis, $p - 1$.
nrbf	the number of RBFs in the model, t.
alpha	a parameter for several RBFs, α.
tol	the tolerance used in the SVD computation, τ.
lambda	the value of the regularisation parameter, λ.
w	the weights, w_k
yhat	the RBF model approximations, $\hat{y}_i$, for $i = 1, 2, \ldots, n$.

Description

Let X be a set of n data records x_i on $p-1$ variables, for $i = 1, 2, \ldots, n$. Furthermore, let the value of the dependent variable for x_i be y_i and assume that:

$$y_i = f(x_i) + \epsilon,$$

for some unknown function $f(\cdot)$ and noise ϵ drawn at random from a Normal distribution with zero mean and unit variance.

A radial basis function (RBF) model approximates the function $f(\cdot)$ by a linear combination of t basis functions, giving an approximate value $\hat{y}_i$ of the dependent value for x_i by calculating:

$$\hat{y}_i = \sum_{k=1}^{t} w_k h_{ik} + b,$$

where b is a scalar intercept term and $h_{ik} = \phi(z_{ik})$ is the value of an RBF $\phi(\cdot)$ for a radial distance z_{ik} between x_i and a centre location c_k of the kth RBF. Equivalently, we can write:

$$\hat{y}_i = \sum_{k=1}^{t+1} w_k h_{ik},$$

by setting $h_{it+1} = 1$, for $i = 1, 2, \ldots, n$.

Radial distances can be computed by one of the following distance functions:

(a) the ℓ_2-norm or Euclidean distance,

$$z_{ik} = \left[(x_i - c_k)S^{-1}(x_i - c_k)^T\right]^{1/2},$$

in which the metric S can be one of:

(i) a diagonal matrix where each of $p-1$ diagonal elements takes a user-supplied value u.
(ii) a diagonal matrix where the jth diagonal element is the standard deviation of the jth independent variable, for $j = 1, 2, \ldots, p-1$.
(iii) the variance-covariance matrix of the data values of independent variables, giving the Mahalanobis distance.
(iv) a diagonal matrix of $p-1$ elements where the jth diagonal element takes the user-supplied value u_j, for $j = 1, 2, \ldots, p-1$.

(b) the ℓ_1-norm or Manhattan distance,

$$z_{ik} = \sum_{j=1}^{p} \left|(x_{ij} - c_{kj})/s_j\right|,$$

where $|\cdot|$ denotes the modulus operator and the scalings s_j can be one of:

(i) a user-supplied value $s_j = u$, for $j = 1, 2, \ldots, p-1$.
(ii) scale by using standard deviations: s_j equals the standard deviation of the jth independent variable, for $j = 1, 2, \ldots, p-1$.
(iii) user-supplied values $s_j = u_j$, for $j = 1, 2, \ldots, p-1$.

The centre locations $\{c_k;\ k = 1, 2, \ldots, t\}$ can be any suitable $(p-1)$-dimensional vectors. Common methods for selecting centre locations of RBFs include:

(a) the values on the independent variables for a random subset of t of the n data records;
(b) the results of a cluster analysis for the independent variables of the n data records for t clusters.

The following radial basis functions $\phi(\cdot)$ are available in NAG DMC:

(a) linear: $\phi(z) = z$;
(b) cubic: $\phi(z) = z^3$;
(c) thin plate spline: $\phi(z) = z^2 \ln(z)$;
(d) Gaussian: $\phi(z) = \exp^{-z^2}$;
(e) multiquadric: $\phi(z, \alpha) = \sqrt{z^2 + \alpha^2}$;
(f) inverse multiquadric: $\phi(z, \alpha) = (z^2 + \alpha^2)^{-1/2}$;
(g) Cauchy: $\phi(z, \alpha) = (z^2 + \alpha^2)^{-1}$.

Values for the weights $w_k;\ k = 1, 2, \ldots, t+1$ are found by minimising the penalised (regularised) sum of squares error function:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{t+1} w_k^2,$$

by computing the pseudo-inverse of a n by $t+1$ matrix H the (i,k)th element of which is h_{ik} and with the value λ added to its diagonal elements. A singular value decomposition is used to compute this inverse with singular values less than the largest singular value multiplied by τ treated as zero; τ can be the value of machine precision or, as suggested by Golub and van Loan (1983), a value consistent with the accuracy of the data.

References and Further Reading

Golub G H and van Loan C F (1983) *Matrix Computations* John Hopkins University Press.

Light W A (1992) Some aspects of radial basis function approximation *Approximation Theory, Spline Functions and Applications* **356** 163–190.

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

Powell M J D (1985) *Radial basis functions for multivariable interpolation: a review* In Cox M G and Mason J C (Editors) *Algorithms for Approximation* Clarendon Press Oxford.

See Also

nagdmc_rbf_predict	predicts values of dependent variables given new data on independent variables and a fitted RBF model.
rbf_ex.c	the example calling program.