NAG Toolbox

nag_smooth_fit_spline (g10ab)

1 Purpose

nag_smooth_fit_spline (g10ab) fits a cubic smoothing spline for a given smoothing parameter.

2 Syntax

```
[yhat, c, rss, df, res, h, comm, ifail] = nag_smooth_fit_spline(mode, x, y, rho, c, comm, 'n', n, 'wt', wt)
[yhat, c, rss, df, res, h, comm, ifail] = g10ab(mode, x, y, rho, c, comm, 'n', n, 'wt', wt)
```

Note: the interface to this routine has changed since earlier releases of the toolbox:

At Mark 24: *weight* was removed from the interface; **wt** was made optional.

3 Description

nag_smooth_fit_spline (g10ab) fits a cubic smoothing spline to a set of n observations $(x_i,\ y_i)$, for $i = 1, 2, \ldots, n$. The spline provides a flexible smooth function for situations in which a simple polynomial or nonlinear regression model is unsuitable.

Cubic smoothing splines arise as the unique real-valued solution function f, with absolutely continuous first derivative and squared-integrable second derivative, which minimizes:

$$\sum_{i=1}^{n} w_i (y_i - f(x_i))^2 + \rho \int_{-\infty}^{\infty} (f''(x))^2 \, dx,$$

where w_i is the (optional) weight for the ith observation and ρ is the smoothing parameter. This criterion consists of two parts: the first measures the fit of the curve, and the second the smoothness of the curve. The value of the smoothing parameter ρ weights these two aspects; larger values of ρ give a smoother fitted curve but, in general, a poorer fit. For details of how the cubic spline can be estimated see Hutchinson and de Hoog (1985) and Reinsch (1967).

The fitted values, $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^{\mathrm{T}}$, and weighted residuals, r_i, can be written as

$$\hat{y} = Hy \quad \text{and} \quad r_i = \sqrt{w_i}(y_i - \hat{y}_i)$$

for a matrix H. The residual degrees of freedom for the spline is trace$(I - H)$ and the diagonal elements of H, h_{ii}, are the leverages.

The parameter ρ can be chosen in a number of ways. The fit can be inspected for a number of different values of ρ. Alternatively the degrees of freedom for the spline, which determines the value of ρ, can be specified, or the (generalized) cross-validation can be minimized to give ρ; see nag_smooth_fit_spline_parest (g10ac) for further details.

nag_smooth_fit_spline (g10ab) requires the x_i to be strictly increasing. If two or more observations have the same x_i-value then they should be replaced by a single observation with y_i equal to the (weighted) mean of the y values and weight, w_i, equal to the sum of the weights. This operation can be performed by nag_smooth_data_order (g10za).

The computation is split into three phases.

(i) Compute matrices needed to fit spline.

(ii) Fit spline for a given value of ρ.

(iii) Compute spline coefficients.

When fitting the spline for several different values of ρ, phase (i) need only be carried out once and then phase (ii) repeated for different values of ρ. If the spline is being fitted as part of an iterative weighted least squares procedure phases (i) and (ii) have to be repeated for each set of weights. In either case, phase (iii) will often only have to be performed after the final fit has been computed.

The algorithm is based on Hutchinson (1986).

4 References

Hastie T J and Tibshirani R J (1990) *Generalized Additive Models* Chapman and Hall

Hutchinson M F (1986) Algorithm 642: A fast procedure for calculating minimum cross-validation cubic smoothing splines *ACM Trans. Math. Software* **12** 150–153

Hutchinson M F and de Hoog F R (1985) Smoothing noisy data with spline functions *Numer. Math.* **47** 99–106

Reinsch C H (1967) Smoothing by spline functions *Numer. Math.* **10** 177–183

5 Parameters

5.1 Compulsory Input Parameters

1: **mode** – CHARACTER(1)

Indicates in which mode the function is to be used.

mode = 'P'
Initialization and fitting is performed. This partial fit can be used in an iterative weighted least squares context where the weights are changing at each call to nag_smooth_fit_spline (g10ab) or when the coefficients are not required.

mode = 'Q'
Fitting only is performed. Initialization must have been performed previously by a call to nag_smooth_fit_spline (g10ab) with **mode** = 'P'. This quick fit may be called repeatedly with different values of **rho** without re-initialization.

mode = 'F'
Initialization and full fitting is performed and the function coefficients are calculated.

Constraint: **mode** = 'P', 'Q' or 'F'.

2: **x**(**n**) – REAL (KIND=nag_wp) array

The distinct and ordered values x_i, for $i = 1, 2, \ldots, n$.

Constraint: $\mathbf{x}(i) < \mathbf{x}(i+1)$, for $i = 1, 2, \ldots, n-1$.

3: **y**(**n**) – REAL (KIND=nag_wp) array

The values y_i, for $i = 1, 2, \ldots, n$.

4: **rho** – REAL (KIND=nag_wp)

ρ, the smoothing parameter.

Constraint: **rho** ≥ 0.0.

5: **c**(ldc, **3**) – REAL (KIND=nag_wp) array

ldc, the first dimension of the array, must satisfy the constraint $ldc \geq \mathbf{n} - 1$.

If **mode** = 'Q', **c** must be unaltered from the previous call to nag_smooth_fit_spline (g10ab) with **mode** = 'P'. Otherwise **c** need not be set.

6: **comm**$(\mathbf{9} \times \mathbf{n} + \mathbf{14})$ – REAL (KIND=nag_wp) array

If **mode** = 'Q', **comm** must be unaltered from the previous call to nag_smooth_fit_spline (g10ab) with **mode** = 'P'. Otherwise **comm** need not be set.

5.2 Optional Input Parameters

1: **n** – INTEGER

Default: the dimension of the arrays **x**, **y**, **comm**. (An error is raised if these dimensions are not equal.)

n, the number of distinct observations.

Constraint: **n** ≥ 3.

2: **wt**$(:)$ – REAL (KIND=nag_wp) array

The dimension of the array **wt** must be at least **n** if $weight$ = 'W'

If $weight$ = 'W', **wt** must contain the n weights. Otherwise **wt** is not referenced and unit weights are assumed.

Constraint: if $weight$ = 'W', **wt**$(i) > 0.0$, for $i = 1, 2, \ldots, n$.

5.3 Output Parameters

1: **yhat**$(\mathbf{n})$ – REAL (KIND=nag_wp) array

The fitted values, $\hat{y}_i$, for $i = 1, 2, \ldots, n$.

2: **c**$(ldc, \mathbf{3})$ – REAL (KIND=nag_wp) array

If **mode** = 'F', **c** contains the spline coefficients. More precisely, the value of the spline at t is given by $((\mathbf{c}(i,3) \times d + \mathbf{c}(i,2)) \times d + \mathbf{c}(i,1)) \times d + \hat{y}_i$, where $x_i \leq t < x_{i+1}$ and $d = t - x_i$.

If **mode** = 'P' or 'Q', **c** contains information that will be used in a subsequent call to nag_smooth_fit_spline (g10ab) with **mode** = 'Q'.

3: **rss** – REAL (KIND=nag_wp)

The (weighted) residual sum of squares.

4: **df** – REAL (KIND=nag_wp)

The residual degrees of freedom.

5: **res**$(\mathbf{n})$ – REAL (KIND=nag_wp) array

The (weighted) residuals, r_i, for $i = 1, 2, \ldots, n$.

6: **h**$(\mathbf{n})$ – REAL (KIND=nag_wp) array

The leverages, h_{ii}, for $i = 1, 2, \ldots, n$.

7: **comm**$(\mathbf{9} \times \mathbf{n} + \mathbf{14})$ – REAL (KIND=nag_wp) array

If **mode** = 'P' or 'Q', **comm** contains information that will be used in a subsequent call to nag_smooth_fit_spline (g10ab) with **mode** = 'Q'.

8: **ifail** – INTEGER

ifail = 0 unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, **n** < 3,
or $ldc <$ **n** $- 1$,
or **rho** < 0.0,
or **mode** $\neq$ 'Q', 'P' or 'F',
or $weight \neq$ 'W' or 'U'.

ifail $= 2$

On entry, $weight =$ 'W' and at least one element of **wt** ≤ 0.0.

ifail $= 3$

On entry, $\mathbf{x}(i) \geq \mathbf{x}(i+1)$, for some i, $i = 1, 2, \ldots, n-1$.

ifail $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

ifail $= -399$

Your licence key may have expired or may not have been installed correctly.

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

Accuracy depends on the value of ρ and the position of the x values. The values of $x_i - x_{i-1}$ and w_i are scaled and ρ is transformed to avoid underflow and overflow problems.

8 Further Comments

The time taken by nag_smooth_fit_spline (g10ab) is of order n.

Regression splines with a small $(< n)$ number of knots can be fitted by nag_fit_1dspline_knots (e02ba) and nag_fit_1dspline_auto (e02be).

9 Example

The data, given by Hastie and Tibshirani (1990), is the age, x_i, and C-peptide concentration (pmol/ml), y_i, from a study of the factors affecting insulin-dependent diabetes mellitus in children. The data is input, reduced to a strictly ordered set by nag_smooth_data_order (g10za) and a series of splines fit using a range of values for the smoothing parameter, ρ.

9.1 Program Text

```
function g10ab_example

fprintf('g10ab example results\n\n');

x =  [ 5.2  8.8 10.5 10.6 10.4  1.8 12.7 15.6  5.8  1.9 ...
       2.2  4.8  7.9  5.2  0.9 11.8  7.9 11.5 10.6  8.5 ...
      11.1 12.8 11.3  1.0 14.5 11.9  8.1 13.8 15.5  9.8 ...
      11.0 12.4 11.1  5.1  4.8  4.2  6.9 13.2  9.9 12.5 ...
      13.2  8.9 10.8];
y =  [ 4.8  4.1  5.2  5.5  5.0  3.4  3.4  4.9  5.6  3.7 ...
       3.9  4.5  4.8  4.9  3.0  4.6  4.8  5.5  4.5  5.3 ...
```

```
       4.7  6.6  5.1  3.9  5.7  5.1  5.2  3.7  4.9  4.8 ...
       4.4  5.2  5.1  4.6  3.9  5.1  5.1  6.0  4.9  4.1 ...
       4.6  4.9  5.1];

% Reorder x, remove ties and weight accordingly
[n, x, y, wt, rss, ifail] = g10za( ...
                                  x, y);
x = x(1:n);
y = y(1:n);

rho  = [1 10 100];
nrho = numel(rho);

c    = zeros(n, 3);
comm = zeros(9*n+14, 1);
yhat = zeros(n,nrho);
rss  = zeros(nrho,1);
df   = zeros(nrho,1);

% Initialize and fit for rho(1)
mode = 'P';
[yhat(:,1), c, rss(1), df(1), res, h, comm, ifail] = ...
  g10ab(mode, x, y, rho(1), c, comm, 'wt', wt);

% Fit for subsequent rhos
mode = 'Q';
for j = 2:nrho
  [yhat(:,j), c, rss(j), df(j), res, h, comm, ifail] = ...
  g10ab( ...
         mode, x, y, rho(j), c, comm, 'wt', wt);
end

%  Display results
fprintf('Smoothing coefficient (rho) = ');
fprintf('  %8.2f', rho);
fprintf('\nResidual sum of squares     = ');
fprintf('%10.3f', rss);
fprintf('\nDegrees of freedom          = ');
fprintf('%10.3f', df);
fprintf('\n\n    x       y                                Fitted Values\n');
fprintf('%8.4f%8.4f%24.4f%10.4f%10.4f\n', [x y yhat]');

fig1 = figure;
plot(x,y,'+',x,yhat(:,1),x,yhat(:,2),x,yhat(:,3));
legend('Raw data', '\rho = 1', '\rho = 10', '\rho = 100', ...
       'Location','NorthWest');
xlabel('Age (years)');
ylabel('C-peptide concentration (pmol/ml)');
title({'Cubic smoothing spline', ...
       'Factors affecting insulin-dependent diabetis mellitus', ...
       'in children; Hastie and Tibshirani (1990)'});
```

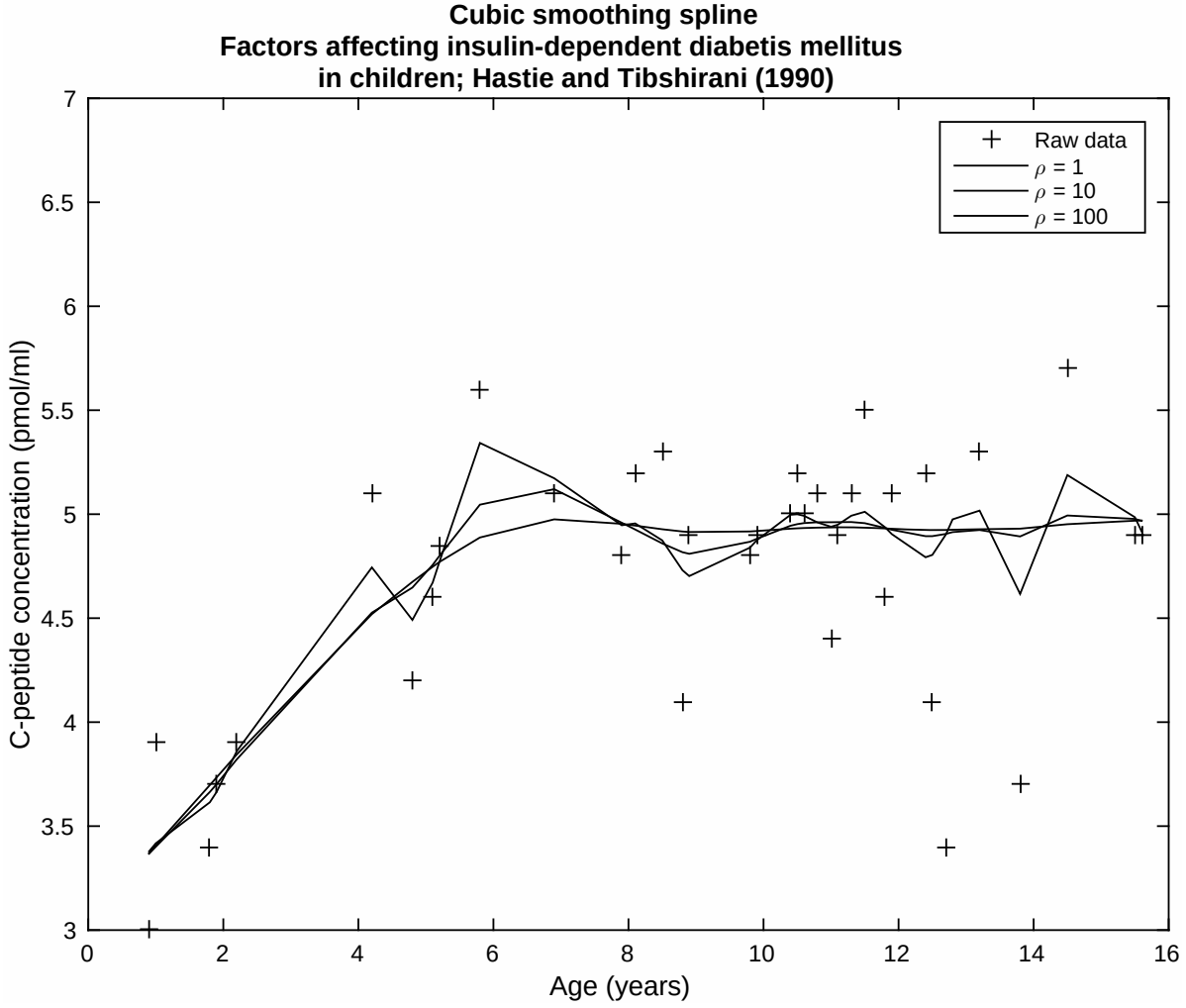
9.2 Program Results

```
     g10ab example results

Smoothing coefficient (rho) =       1.00     10.00    100.00
Residual sum of squares     =      9.118    11.288    11.881
Degrees of freedom          =     22.505    27.785    31.191

    x       y                          Fitted Values
  0.9000  3.0000                 3.3784    3.3674    3.3699
  1.0000  3.9000                 3.4173    3.4008    3.4063
  1.8000  3.4000                 3.6144    3.6642    3.6973
  1.9000  3.7000                 3.6639    3.7016    3.7341
  2.2000  3.9000                 3.8607    3.8214    3.8449
  4.2000  5.1000                 4.7441    4.5265    4.5194
  4.8000  4.2000                 4.4914    4.6471    4.6746
  5.1000  4.6000                 4.6708    4.7561    4.7470
  5.2000  4.8500                 4.7704    4.7993    4.7702
```

5.8000	5.6000	5.3426	5.0458	4.8879
6.9000	5.1000	5.1728	5.1204	4.9753
7.9000	4.8000	4.9467	4.9590	4.9537
8.1000	5.2000	4.9556	4.9262	4.9452
8.5000	5.3000	4.8742	4.8595	4.9276
8.8000	4.1000	4.7305	4.8172	4.9168
8.9000	4.9000	4.7024	4.8095	4.9143
9.8000	4.8000	4.8394	4.8676	4.9170
9.9000	4.9000	4.8746	4.8818	4.9191
10.4000	5.0000	4.9971	4.9445	4.9303
10.5000	5.2000	4.9997	4.9521	4.9321
10.6000	5.0000	4.9921	4.9572	4.9335
10.8000	5.1000	4.9603	4.9613	4.9354
11.0000	4.4000	4.9396	4.9614	4.9363
11.1000	4.9000	4.9494	4.9618	4.9366
11.3000	5.1000	4.9926	4.9623	4.9366
11.5000	5.5000	5.0116	4.9568	4.9355
11.8000	4.6000	4.9372	4.9338	4.9315
11.9000	5.1000	4.9042	4.9251	4.9300
12.4000	5.2000	4.7929	4.8943	4.9240
12.5000	4.1000	4.8042	4.8944	4.9237
12.7000	3.4000	4.9020	4.9051	4.9244
12.8000	6.6000	4.9752	4.9138	4.9252
13.2000	5.3000	5.0173	4.9239	4.9276
13.8000	3.7000	4.6164	4.8930	4.9304
14.5000	5.7000	5.1883	4.9938	4.9518
15.5000	4.9000	4.9854	4.9773	4.9687
15.6000	4.9000	4.9167	4.9682	4.9697


Cubic smoothing spline
Factors affecting insulin-dependent diabetis mellitus
in children; Hastie and Tibshirani (1990)
Raw data
ρ = 1
ρ = 10
ρ = 100
C-peptide concentration (pmol/ml)
Age (years)
7
6.5
6
5.5
5
4.5
4
3.5
3
0
2
4
6
8
10
12
14
16