

NAG Library Routine Document

G02KBF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G02KBF calculates a ridge regression, with ridge parameters supplied by you.

2 Specification

```

SUBROUTINE G02KBF(N, M, X, LDX, ISX, IP, Y, LH, H, NEP, WANTB, B, LDB,
1          WANTVF, VF, LDVF, LPEC, PEC, PE, LDPE, IFAIL)
    INTEGER          N, M, LDX, ISX(M), IP, LH, WANTB, LDB, WANTVF, LDVF,
1          LPEC, LDPE, IFAIL
    double precision X(LDX,M), Y(N), H(LH), NEP(LH), B(LDB,*), VF(LDVF,*),
1          PE(LDPE,*),
    CHARACTER*1      PEC(LPEC)

```

3 Description

A linear model has the form:

$$y = c + X\beta + \epsilon,$$

where

y is an n by 1 matrix of values of a dependent variable;

c is a scalar intercept term;

X is an n by m matrix of values of independent variables;

β is a m by 1 matrix of unknown values of parameters;

ϵ is an n by 1 matrix of unknown random errors such that variance of $\epsilon = \sigma^2 I$.

Let $\tilde{X}$ be the mean-centred X and $\tilde{y}$ the mean-centred y . Furthermore, $\tilde{X}$ is scaled such that the diagonal elements of the cross product matrix $\tilde{X}^T \tilde{X}$ are one. The linear model now takes the form:

$$\tilde{y} = \tilde{X}\tilde{\beta} + \epsilon.$$

Ridge regression estimates the parameters $\tilde{\beta}$ in a penalised least squares sense by finding the $\tilde{b}$ that minimizes

$$\|\tilde{X}\tilde{b} - \tilde{y}\|^2 + h\|\tilde{b}\|^2, \quad h > 0,$$

where $\|\cdot\|$ denotes the ℓ_2 -norm and h is a scalar regularization or ridge parameter. For a given value of h , the parameters estimates $\tilde{b}$ are found by evaluating

$$\tilde{b} = (\tilde{X}^T \tilde{X} + hI)^{-1} \tilde{X}^T \tilde{y}.$$

Note that if $h = 0$ the ridge regression solution is equivalent to the ordinary least squares solution.

Rather than calculate the inverse of $(\tilde{X}^T \tilde{X} + hI)$ directly, G02KBF uses the singular value decomposition (SVD) of $\tilde{X}$. After decomposing $\tilde{X}$ into UDV^T where U and V are orthogonal matrices and D is a diagonal matrix, the parameter estimates become

$$\tilde{b} = V(D^T D + hI)^{-1} D U^T \tilde{y}.$$

A consequence of introducing the ridge parameter is that the effective number of parameters, γ , in the model is given by the sum of diagonal elements of

$$D^T D (D^T D + hI)^{-1},$$

see Moody (1992) for details.

Any multi-collinearity in the design matrix X may be highlighted by calculating the variance inflation factors for the fitted model. The j th variance inflation factor, v_j , is a scaled version of the multiple correlation coefficient between independent variable j and the other independent variables, R_j , and is given by

$$v_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \dots, m.$$

The m variance inflation factors are calculated as the diagonal elements of the matrix:

$$(\tilde{X}^T \tilde{X} + hI)^{-1} \tilde{X}^T \tilde{X} (\tilde{X}^T \tilde{X} + hI)^{-1},$$

which, using the SVD of $\tilde{X}$, is equivalent to the diagonal elements of the matrix:

$$V(D^T D + hI)^{-1} D^T D (D^T D + hI)^{-1} V^T.$$

Given a value of h , any or all of the following prediction criteria are available:

(a) Generalized cross-validation (GCV):

$$\frac{ns}{(n - \gamma)^2};$$

(b) Unbiased estimate of variance (UEV):

$$\frac{s}{n - \gamma};$$

(c) Future prediction error (FPE):

$$\frac{1}{n} \left(s + \frac{2\gamma s}{n - \gamma} \right);$$

(d) Bayesian information criterion (BIC):

$$\frac{1}{n} \left(s + \frac{\log(n)\gamma s}{n - \gamma} \right);$$

(e) Leave-one-out cross-validation (LOOCV),

where s is the sum of squares of residuals.

Although parameter estimates $\tilde{b}$ are calculated by using $\tilde{X}$, it is usual to report the parameter estimates b associated with X . These are calculated from $\tilde{b}$, and the means and scalings of X . Optionally, either $\tilde{b}$ or b may be calculated.

4 References

Hastie T, Tibshirani R and Friedman J (2003) *The Elements of Statistical Learning: Data Mining, Inference and Prediction* Springer Series in Statistics

Moody J.E. (1992) The effective number of parameters: An analysis of generalisation and regularisation in nonlinear learning systems *In: Neural Information Processing Systems* (eds J E Moody, S J Hanson, and R P Lippmann) 4 847–854 Morgan Kaufmann San Mateo CA

5 Parameters

- 1: N – INTEGER *Input*
On entry: n , the number of observations.
Constraint: $N \geq 1$.

- 2: M – INTEGER *Input*
On entry: the number of independent variables available in the data matrix X .
Constraint: $M \leq N$.

- 3: X(LDX,M) – **double precision** array *Input*
On entry: the values of independent variables in the data matrix X .

- 4: LDX – INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G02KBF is called.
Constraint: $LDX \geq N$.

- 5: ISX(M) – INTEGER array *Input*
On entry: indicates which m independent variables are included in the model.
 $ISX(j) = 1$
 The j th variable in X will be included in the model.
 $ISX(j) = 0$
 Variable j is excluded.
Constraint: $ISX(j) = 0$ or 1 , for $j = 1, 2, \dots, M$.

- 6: IP – INTEGER *Input*
On entry: m , the number of independent variables in the model.
Constraints:
 $1 \leq IP \leq M$;
 exactly IP elements of ISX must be equal to 1 .

- 7: Y(N) – **double precision** array *Input*
On entry: the n values of the dependent variable y .

- 8: LH – INTEGER *Input*
On entry: the number of supplied ridge parameters.
Constraint: $LH > 0$.

- 9: H(LH) – **double precision** array *Input*
On entry: $H(j)$ is the value of the j th ridge parameter h .
Constraint: $H(j) \geq 0.0$, for $j = 1, 2, \dots, LH$.

- 10: NEP(LH) – **double precision** array *Output*
On exit: $NEP(j)$ is the number of effective parameters, γ , in the j th model, for $j = 1, 2, \dots, LH$.

- 11: WANTB – INTEGER *Input*
- On entry:* defines the options for parameter estimates.
- WANTB = 0
Parameter estimates are not calculated and B is not referenced.
- WANTB = 1
Parameter estimates b are calculated for the original data.
- WANTB = 2
Parameter estimates $\tilde{b}$ are calculated for the standardized data.
- Constraint:* WANTB = 0, 1 or 2.
- 12: B(LDB,*) – **double precision** array *Output*
- Note:** the second dimension of the array B must be at least LH if WANTB $\neq$ 0, and at least 1 otherwise.
- On exit:* if WANTB $\neq$ 0, B contains the intercept and parameter estimates for the fitted ridge regression model in the order indicated by ISX. B(1, j), for $j = 1, 2, \dots, \text{LH}$, contains the estimate for the intercept; B($i + 1, j$) contains the parameter estimate for the i th independent variable in the model fitted with ridge parameter H(j), for $i = 1, 2, \dots, \text{IP}$.
- 13: LDB – INTEGER *Input*
- On entry:* the first dimension of the array B as declared in the (sub)program from which G02KBF is called.
- Constraints:*
- if WANTB $\neq$ 0, LDB $\geq$ IP + 1;
otherwise LDB $\geq$ 1.
- 14: WANTVF – INTEGER *Input*
- On entry:* defines the options for variance inflation factors.
- WANTVF = 0
Variance inflation factors are not calculated and the array VF is not referenced.
- WANTVF = 1
Variance inflation factors are calculated.
- Constraints:*
- WANTVF = 0 or 1;
if WANTB = 0, WANTVF $\neq$ 0.
- 15: VF(LDVF,*) – **double precision** array *Output*
- Note:** the second dimension of the array VF must be at least LH if WANTVF $\neq$ 0, and at least 1 otherwise.
- On exit:* if WANTVF = 1, the variance inflation factors. For the i th independent variable in a model fitted with ridge parameter H(j), VF(i, j) is the value of v_i , for $i = 1, 2, \dots, \text{IP}$.
- 16: LDVF – INTEGER *Input*
- On entry:* the first dimension of the array VF as declared in the (sub)program from which G02KBF is called.
- Constraints:*
- if WANTVF $\neq$ 0, LDVF $\geq$ IP;
otherwise LDVF $\geq$ 1.

- 17: LPEC – INTEGER *Input*
On entry: the number of prediction error statistics to return; set $LPEC \leq 0$ for no prediction error estimates.
- 18: PEC(LPEC) – CHARACTER*1 array *Input*
On entry: if $LPEC > 0$, $PEC(j)$ defines the j th prediction error, for $j = 1, 2, \dots, LPEC$; otherwise PEC is not referenced.
 $PEC(j) = 'B'$
 Bayesian information criterion (BIC).
 $PEC(j) = 'F'$
 Future prediction error (FPE).
 $PEC(j) = 'G'$
 Generalized cross-validation (GCV).
 $PEC(j) = 'L'$
 Leave-one-out cross-validation (LOOCV).
 $PEC(j) = 'U'$
 Unbiased estimate of variance (UEV).
Constraint: if $LPEC > 0$, $PEC(j) = 'B', 'F', 'G', 'L'$ or $'U'$ for $j = 1, 2, \dots, LPEC$.
- 19: PE(LDPE,*) – **double precision** array *Output*
Note: the second dimension of the array PE must be at least LH if $LPEC > 0$, and at least 1 otherwise.
On exit: if $LPEC \leq 0$, PE is not referenced; otherwise $PE(i, j)$ contains the prediction error of criterion $PEC(i)$ for the model fitted with ridge parameter $H(j)$, for $i = 1, 2, \dots, LPEC$ and $j = 1, 2, \dots, LH$.
- 20: LDPE – INTEGER *Input*
On entry: the first dimension of the array PE as declared in the (sub)program from which G02KBF is called.
Constraints:
 if $LPEC > 0$, $LDPE \geq LPEC$;
 otherwise $LDPE \geq 1$.
- 21: IFAIL – INTEGER *Input/Output*
On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.
On exit: $IFAIL = 0$ unless the routine detects an error (see Section 6).
 For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, if you are not familiar with this parameter, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry $IFAIL = 0$ or -1 , explanatory error messages are output on the current error message unit (as defined by $X04AAF$).

Errors or warnings detected by the routine:

$IFAIL = 1$

On entry, $N \leq 1$,
 or $H(j) < 0.0$,
 or $LH \leq 0$,
 or $WANTB \neq 0, 1$ or 2 ,
 or $WANTB \neq 0$ and $LDB < IP + 1$,
 or $WANTVF \neq 0$ or 1 ,
 or an element of PEC is not defined.

$IFAIL = 2$

On entry, $M > N$,
 or $LDX < N$,
 or $IP < 1$ or $IP > M$,
 or an element of $ISX \neq 0$ or 1 ,
 or IP does not equal the sum of elements in ISX ,
 or $WANTVF \neq 0$ and $LDVF < IP$,
 or $LDPE < LPEC$.

$IFAIL = 3$

Both $WANTB$ and $WANTVF$ are zero.

$IFAIL = 4$

Internal error. Check all array sizes and calls to $G02KBF$. Please contact NAG.

$IFAIL = -999$

Internal memory allocation failed.

7 Accuracy

The accuracy of $G02KBF$ is closely related to that of the singular value decomposition.

8 Further Comments

$G02KBF$ allocates internally $\max(5 \times (N - 1), 2 \times IP \times IP) + (N + 3) \times IP + N$ elements of double precision storage.

9 Example

This example reads in data from an experiment to model body fat, and a selection of ridge regression models are calculated.

9.1 Program Text

```
*      G02KBF Example Program Text
*      Mark 22 Release. NAG Copyright 2008.
*      .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          LDX, LDB, HMAX, LDPE
      PARAMETER        (LDX=50,LDB=10,HMAX=20,LDPE=5)
      INTEGER          LDVF
```

```

      PARAMETER          (LDVF=LDB-1)
*   .. Local Scalars ..
      INTEGER            I, IFAIL, IP, IP1, J, LH, LPEC, M, N, PL, WANTB,
+
      WANTVF
*   .. Local Arrays ..
      DOUBLE PRECISION  B(LDB,HMAX), H(HMAX), NEP(HMAX), PE(LDPE,HMAX),
+
      VF(LDVF,HMAX), X(LDX,LDB), Y(LDX)
      INTEGER            ISX(HMAX)
      CHARACTER          PEC(LDPE)
*   .. External Subroutines ..
      EXTERNAL          G02KBF
*   .. Intrinsic Functions ..
      INTRINSIC          MIN
*   .. Executable Statements ..
      CONTINUE

*
      WRITE (NOUT,*) 'G02KBF Example Program Results'
*
*   Skip heading in data file
      READ (NIN,*)
*
*   Read in the problem size information
      READ (NIN,*) N, M, LH, LPEC, WANTB
*
*   Check array sizes
      IF (M.GT.LDB .OR. N.GT.LDX .OR. LH.GT.HMAX .OR. LPEC.GT.LDPE) THEN
          WRITE (NOUT,99990)
          GO TO 160
      END IF
*
*   Read in the data
      IF (LPEC.GT.0) THEN
          READ (NIN,*) (PEC(I),I=1,LPEC)
      END IF
      DO 20 I = 1, N
          READ (NIN,*) (X(I,J),J=1,M), Y(I)
20  CONTINUE
*
*   Read in the ISX flags
      READ (NIN,*) (ISX(I),I=1,M)
*
*   Read in the ridge coefficients
      READ (NIN,*) (H(I),I=1,LH)
*
*   Total number of variables
      IP = 0
      DO 40 J = 1, M
          IF (ISX(J).EQ.1) THEN
              IP = IP + 1
          END IF
40  CONTINUE
*
*   Output the variance inflation factors and parameter estimates
*   (original scalings)
      WANTVF = 1
*
*   Run the analysis
      IFAIL = 1
      CALL G02KBF(N,M,X,LDX,ISX,IP,Y,LH,H,NEP,WANTB,B,LDB,WANTVF,VF,
+
      LDVF,LPEC,PEC,PE,LDPE,IFAIL)
      IF (IFAIL.NE.0) THEN
          WRITE (NOUT,99991) IFAIL
          GO TO 160
      END IF
*   Output results
      IP1 = IP - 1
*
*   Summaries
      WRITE (NOUT,99994) 'Number of parameters used = ', IP + 1
      WRITE (NOUT,*) 'Effective number of parameters (NEP):'
      WRITE (NOUT,*) '   Ridge   '

```

```

WRITE (NOUT,*) '    Coeff.  ', 'NEP'
DO 60 I = 1, LH
  WRITE (NOUT,99993) H(I), NEP(I)
60 CONTINUE
*
*   Parameter estimates
IF (WANTB.NE.0) THEN
  WRITE (NOUT,*)
  IF (WANTB.EQ.1) THEN
    WRITE (NOUT,*) 'Parameter Estimates (Original scalings)'
  ELSE
    WRITE (NOUT,*) 'Parameter Estimates (Standarised)'
  END IF
  PL = MIN(IP,4)
  WRITE (NOUT,*) '    Ridge  '
  WRITE (NOUT,99997) '    Coeff.  ', ' Intercept ', (I,I=1,PL)
  IF (PL.LT.IP1) THEN
    WRITE (NOUT,99996) (I,I=PL+1,IP1)
  END IF
  PL = MIN(IP+1,5)
  DO 80 I = 1, LH
    WRITE (NOUT,99999) H(I), (B(J,I),J=1,PL)
    IF (PL.LT.IP) THEN
      WRITE (NOUT,99998) (B(J,I),J=PL+1,IP)
    END IF
  80 CONTINUE
  END IF
*
*   Variance inflation factors
IF (WANTVF.NE.0) THEN
  WRITE (NOUT,*)
  WRITE (NOUT,*) 'Variance Inflation Factors'
  PL = MIN(IP,5)
  WRITE (NOUT,*) '    Ridge  '
  WRITE (NOUT,99995) '    Coeff.  ', (I,I=1,PL)
  IF (PL.LT.IP) THEN
    WRITE (NOUT,99996) (I,I=PL+1,IP)
  END IF
  DO 100 I = 1, LH
    WRITE (NOUT,99999) H(I), (VF(J,I),J=1,PL)
    IF (PL.LT.IP) THEN
      WRITE (NOUT,99998) (VF(J,I),J=PL+1,IP)
    END IF
  100 CONTINUE
  END IF
*
*   Prediction error criterion
IF (LPEC.GT.0) THEN
  WRITE (NOUT,*)
  WRITE (NOUT,*) 'Prediction error criterion'
  PL = MIN(LPEC,5)
  WRITE (NOUT,*) '    Ridge  '
  WRITE (NOUT,99995) '    Coeff.  ', (I,I=1,PL)
  IF (PL.LT.LPEC) THEN
    WRITE (NOUT,99996) (I,I=PL+1,LPEC)
  END IF
  DO 120 I = 1, LH
    WRITE (NOUT,99999) H(I), (PE(J,I),J=1,PL)
    IF (PL.LT.IP) THEN
      WRITE (NOUT,99998) (PE(J,I),J=PL+1,IP)
    END IF
  120 CONTINUE
  WRITE (NOUT,*)
  WRITE (NOUT,*) 'Key:'
  DO 140 I = 1, LPEC
    IF (PEC(I).EQ.'L') THEN
      WRITE (NOUT,99992) I, 'Leave one out cross-validation'
    ELSE IF (PEC(I).EQ.'G') THEN
      WRITE (NOUT,99992) I, 'Generalised cross-validation'
    ELSE IF (PEC(I).EQ.'U') THEN
      WRITE (NOUT,99992) I, 'Unbiased estimate of variance'
    
```

```

        ELSE IF (PEC(I).EQ.'F') THEN
            WRITE (NOUT,99992) I, 'Final prediction error'
        ELSE IF (PEC(I).EQ.'B') THEN
            WRITE (NOUT,99992) I, 'Bayesian information criterion'
        END IF
140     CONTINUE
        END IF
*
160 CONTINUE
*
99999 FORMAT (F10.4,5F10.4)
99998 FORMAT (10X,5F10.4)
99997 FORMAT (A,A,4I10)
99996 FORMAT (10X,5I10)
99995 FORMAT (A,5I10)
99994 FORMAT (A,I10)
99993 FORMAT (F10.4,F10.4)
99992 FORMAT (1X,I5,1X,A)
99991 FORMAT (1X,/1X,' ** G02KBF returned with IFAIL = ',I5)
99990 FORMAT (1X,' ** Problem size too large, increase array limits')
END

```

9.2 Program Data

G02KBF Example Program Data

```

20   3   16   5   1       : N, M, LH, LPEC, WANTB
L G U F B : PEC
19.5  43.1  29.1  11.9
24.7  49.8  28.2  22.8
30.7  51.9  37.0  18.7
29.8  54.3  31.1  20.1
19.1  42.2  30.9  12.9
25.6  53.9  23.7  21.7
31.4  58.5  27.6  27.1
27.9  52.1  30.6  25.4
22.1  49.9  23.2  21.3
25.5  53.5  24.8  19.3
31.1  56.6  30.0  25.4
30.4  56.7  28.3  27.2
18.7  46.5  23.0  11.7
19.7  44.2  28.6  17.8
14.6  42.7  21.3  12.8
29.5  54.4  30.1  23.9
27.7  55.3  25.7  22.6
30.2  58.6  24.6  25.4
22.7  48.2  27.1  14.8
25.2  51.0  27.5  21.1 : End of observations
1      1      1      : ISX
0.0    0.002  0.004  0.006
0.008  0.010  0.012  0.014
0.016  0.018  0.020  0.022
0.024  0.026  0.028  0.030 : Ridge co-efficients

```

9.3 Program Results

G02KBF Example Program Results

Number of parameters used = 4

Effective number of parameters (NEP):

Ridge	Coeff.	NEP
0.0000	4.0000	
0.0020	3.2634	
0.0040	3.1475	
0.0060	3.0987	
0.0080	3.0709	
0.0100	3.0523	
0.0120	3.0386	
0.0140	3.0278	
0.0160	3.0189	
0.0180	3.0112	

0.0200	3.0045
0.0220	2.9984
0.0240	2.9928
0.0260	2.9876
0.0280	2.9828
0.0300	2.9782

Parameter Estimates (Original scalings)

Ridge	Intercept	1	2	3
0.0000	117.0847	4.3341	-2.8568	-2.1861
0.0020	22.2748	1.4644	-0.4012	-0.6738
0.0040	7.7209	1.0229	-0.0242	-0.4408
0.0060	1.8363	0.8437	0.1282	-0.3460
0.0080	-1.3396	0.7465	0.2105	-0.2944
0.0100	-3.3219	0.6853	0.2618	-0.2619
0.0120	-4.6734	0.6432	0.2968	-0.2393
0.0140	-5.6511	0.6125	0.3222	-0.2228
0.0160	-6.3891	0.5890	0.3413	-0.2100
0.0180	-6.9642	0.5704	0.3562	-0.1999
0.0200	-7.4236	0.5554	0.3681	-0.1916
0.0220	-7.7978	0.5429	0.3779	-0.1847
0.0240	-8.1075	0.5323	0.3859	-0.1788
0.0260	-8.3673	0.5233	0.3926	-0.1737
0.0280	-8.5874	0.5155	0.3984	-0.1693
0.0300	-8.7758	0.5086	0.4033	-0.1653

Variance Inflation Factors

Ridge	1	2	3
0.0000	708.8429	564.3434	104.6060
0.0020	50.5592	40.4483	8.2797
0.0040	16.9816	13.7247	3.3628
0.0060	8.5033	6.9764	2.1185
0.0080	5.1472	4.3046	1.6238
0.0100	3.4855	2.9813	1.3770
0.0120	2.5434	2.2306	1.2356
0.0140	1.9581	1.7640	1.1463
0.0160	1.5698	1.4541	1.0859
0.0180	1.2990	1.2377	1.0428
0.0200	1.1026	1.0805	1.0105
0.0220	0.9556	0.9627	0.9855
0.0240	0.8427	0.8721	0.9655
0.0260	0.7541	0.8007	0.9491
0.0280	0.6832	0.7435	0.9353
0.0300	0.6257	0.6969	0.9235

Prediction error criterion

Ridge	1	2	3	4	5
0.0000	8.0368	7.6879	6.1503	7.3804	8.6052
0.0020	7.5464	7.4238	6.2124	7.2261	8.2355
0.0040	7.5575	7.4520	6.2793	7.2675	8.2515
0.0060	7.5656	7.4668	6.3100	7.2876	8.2611
0.0080	7.5701	7.4749	6.3272	7.2987	8.2661
0.0100	7.5723	7.4796	6.3381	7.3053	8.2685
0.0120	7.5732	7.4823	6.3455	7.3095	8.2695
0.0140	7.5734	7.4838	6.3508	7.3122	8.2696
0.0160	7.5731	7.4845	6.3548	7.3140	8.2691
0.0180	7.5724	7.4848	6.3578	7.3151	8.2683
0.0200	7.5715	7.4847	6.3603	7.3158	8.2671
0.0220	7.5705	7.4843	6.3623	7.3161	8.2659
0.0240	7.5694	7.4838	6.3639	7.3162	8.2645
0.0260	7.5682	7.4832	6.3654	7.3162	8.2630
0.0280	7.5669	7.4825	6.3666	7.3161	8.2615
0.0300	7.5657	7.4818	6.3677	7.3159	8.2600

Key:

- 1 Leave one out cross-validation
 - 2 Generalised cross-validation
 - 3 Unbiased estimate of variance
 - 4 Final prediction error
 - 5 Bayesian information criterion
-