Algorithm AS 148

The Jackknife

By Norman S. Matloff

Division of Statistics, University of California at Davis, Davis, CA, U.S.A.

Keywords: JACKKNIFE

Language

ISO Fortran

Description and Purpose

Suppose $\hat{\theta}$ is an estimate of a parameter θ, based on a random sample $X_1, \ldots, X_n$; here we allow both θ and the X_i to be vector-valued. For each $i = 1, \ldots, n$, define the "pseudovalue" $\hat{\theta}_{J,i}$ to be $n\hat{\theta} - (n-1)\hat{\theta}_{(-i)}$, where $\hat{\theta}_{(-i)}$ is the estimate of θ which has the same form as $\hat{\theta}$ but which is based on $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. Then $\hat{\theta}_J = n^{-1} \Sigma \hat{\theta}_{J,i}$, summing over i from 1 to n, is called the jackknife estimator of θ corresponding to $\hat{\theta}$.

In the "smooth" case, $\hat{\theta}_J$ has smaller bias than $\hat{\theta}$ and is asymptotically normally distributed with estimated covariance matrix $n^{-1}\{(n-1)^{-1}(\hat{\theta}_{J,i} - \hat{\theta}_J)(\hat{\theta}_{J,i} - \hat{\theta}_J)^t\}$ (see Cox and Hinkley, 1975). The second of these two properties is useful in several ways:

1. $\hat{\theta}_J$ can be used to form distribution-free confidence intervals and hypothesis tests. For example, the classical χ^2-based inference procedures for a population variance σ^2 are not robust with respect to the assumption that the sampled population is normally distributed. However, we can apply the jackknife procedure to the sample variance s^2 (or for greater accuracy, to $\log(s^2)$) to derive large-sample inference procedures which are valid without the normality assumption.
2. $\hat{\theta}_J$ can also be used to make some inferences more convenient. For example, suppose $\hat{\theta} = f(\hat{\gamma}_1, \ldots, \hat{\gamma}_r)$, where f is a smooth, real-valued function and $(\hat{\gamma}_1, \ldots, \hat{\gamma}_r)^t$ is an asymptotically multivariate normally distributed estimator of $(\gamma_1, \ldots, \gamma_r)^t$. Here we might consider making inferences on $\theta = f(\gamma_1, \ldots, \gamma_r)$ using the "delta method" (Rao, 1973, p. 388). However, this may be inconvenient if the derivatives of f are complicated. Use of the jackknife procedure (with the algorithm presented here) may be much more convenient.
3. The pseudovalues $\hat{\theta}_{J,i}$ may be useful in detecting outliers and evaluating their influence on $\hat{\theta}$ and $\hat{\theta}_J$.

In some cases, the sample size n may be too large for economical computation. Here one can use "jackknifing by groups": Suppose $n = gl$ for positive integers g and l. Divide the sample into g groups of size l. In this version of the jackknife, $\hat{\theta}_{(-i)}$ and $\hat{\theta}_{J,i}$ are based on deletion of the ith group of observations from the sample, rather than deletion of the ith observation. Also, in the expressions for $\hat{\theta}_{J,i}$, $\hat{\theta}_J$ and the estimated covariance matrix of $\hat{\theta}_J$, the quantity n is replaced everywhere by g.

A nice review paper on the jackknife is Miller (1974).

Structure

SUBROUTINE JKKN(X, K, N, TH, THJ, LTH, COV, PS, NGRP, IGSIZE, IFAULT)

Formal parameters

X	Real array (K, N)	input: contains the sample $X_1, \ldots, X_N$; initially $X(J, I)$ contains the Jth component of X_I; the array is rearranged during computation
K		input: number of components in each observation
N		input: sample size

TH	Real array (*LTH*)	output :	work array
THJ	Real array (*LTH*)	output :	$\hat{\theta}_J$
LTH		input :	number of components in θ
COV	Real array (*LTH*, *LTH*)	output :	estimated covariance matrix of $\hat{\theta}_J$
PS	Real array (*LTH*, *NGRP*)	output :	contains the pseudovalues $\hat{\theta}_{J,1}, \ldots, \hat{\theta}_{J,NGRP}$; $PS(L,I)$ contains the Lth component of $\hat{\theta}_{J,L}$
NGRP		input :	g, the number of groups (ordinarily equal to N)
IGSIZE		input :	l, the group size (ordinarily equal to 1)
IFAULT		output :	equals 1 if $NGRP = 1$ (in which case computation is suppressed); equals 0 otherwise

Auxiliary subroutines:

The base estimator $\hat{\theta}$ is calculated by the user-supplied subroutine *THTHT*, whose form should be

SUBROUTINE THTHT(X, K, M, TH, LTH)
REAL X(K,M), TH(LTH)

[Here $\hat{\theta}$ is calculated, based on the first M observations currently stored in X. The array X is assumed to have the same structure as in *JKKN*, although *JKKN* usually calls *THTHT* with M equal to $N - IGSIZE$ rather than N. The value $\hat{\theta}$ is returned as an output to *JKKN* through the *THTHT* array *TH*; all other parameters in *THTHT* are inputs to that routine.]

RETURN
END

References

Cox, D. R. and Hinkley, D. V. (1975). *Theoretical Statistics.* London : Chapman and Hall.
Miller, R. G. (1974). The jackknife—a review. *Biometrika*, **61**, 1–15.
Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. London : Wiley.

```
      SUBROUTINE JKKN(X, K, N, TH, THJ, LTH, COV, PS, NGRP,
     *  IGSIZE, IFAULT)
C
C        ALGORITHM AS 148  APPL. STATIST. (1980) VOL.29, NO.1
C
C        REMOVAL OF BIAS BY THE JACKKNIFE PROCEDURE
C
      REAL X(K, N), TH(LTH), THJ(LTH), COV(LTH, LTH), PS(LTH, NGRP)
      IFAULT = 1
      IF (NGRP .LE. 1) RETURN
      IFAULT = 0
      NGRP1 = NGRP - 1
      NN = NGRP1 * IGSIZE
      ENGRP = NGRP
      ENGRP1 = NGRP1
      ENN = ENGRP * ENGRP1
      IGSZ1 = IGSIZE + 1
      KI = K * IGSIZE
C
C        KI = NUMBER OF X COMPONENTS IN A GROUP.
C
      LLN = NGRP1 * KI
C
C        FIRST CALCULATE NGRP * THETAHAT AND STORE IT IN TH.
C
      CALL THTHT(X, K, N, TH, LTH)
      DO 10 I = 1, LTH
   10 TH(I) = ENGRP * TH(I)
C
C        CALCULATE THE NGRP PSEUDOVALUES.
C
```

```
      IG1 = 0
      DO 50 I = 1, NGRP
      IF (I .EQ. NGRP) GOTO 25
C
C        TRADE THE I-TH GROUP OF OBSERVATIONS
C        (AS STORED CURRENTLY) WITH THE NGRP-TH
C        GROUP, SO AS TO DELETE THE I-TH GROUP.
C
      IG2 = NN
      DO 20 IG = 1, IGSIZE
      IG1 = IG1 + 1
      IG2 = IG2 + 1
      DO 15 J = 1, K
      TEMP = X(J, IG2)
      X(J, IG2) = X(J, IG1)
      X(J, IG1) = TEMP
   15 CONTINUE
   20 CONTINUE
      IF (I .LT. NGRP) CALL THTHT(X, K, NN, THJ, LTH)
   25 IF (I .EQ. NGRP) CALL THTHT(X(1, IGSZ1), K, NN, THJ, LTH)
      DO 35 II = 1, LTH
   35 PS(II, I) = TH(II) - THJ(II) * ENGRP1
   50 CONTINUE
C
C        INITIALIZE.
C
      DO 75 I = 1, LTH
      THJ(I) = 0.0
      DO 75 J = 1, LTH
      COV(I, J) = 0.0
   75 CONTINUE
C
C        CALCULATE JACKKNIFED VERSION OF THETAHAT
C        AND STORE IT IN THJ.
C
      DO 100 I = 1, NGRP
      DO 90 II = 1, LTH
   90 THJ(II) = THJ(II) + PS(II, I)
  100 CONTINUE
      DO 110 II = 1, LTH
  110 THJ(II) = THJ(II) / ENGRP
C
C        CALCULATE THE APPROXIMATE COVARIANCE MATRIX OF
C        THE JACKKNIFED THETAHAT, AND STORE IT IN COV.
C
      DO 125 II = 1, LTH
      THJII = THJ(II)
      DO 125 JJ = II, LTH
      THJJJ = THJ(JJ)
      DO 115 I = 1, NGRP
      COV(II, JJ) = COV(II, JJ) +
     *  (PS(II, I) - THJII) * (PS(JJ, I) - THJJJ)
  115 CONTINUE
  125 CONTINUE
C
C        IF LTH .GT. 1, COPY THE UPPER-TRIANGULAR SECTION
C        OF COV INTO THE LOWER-TRIANGULAR SECTION.
C
      IF (LTH .EQ. 1) GOTO 200
      DO 175 II = 2, LTH
      II1 = II - 1
      DO 175 JJ = 1, II1
      COV(II, JJ) = COV(JJ, II)
  175 CONTINUE
  200 DO 210 II = 1, LTH
      DO 210 JJ = 1, LTH
  210 COV(II, JJ) = COV(II, JJ) / ENN
      RETURN
      END
```