

AD-A119 659

RENSSELAER POLYTECHNIC INST TROY NY

F/6 12/1

SELF-CRITICAL, AND ROBUST, ESTIMATES FOR THE PARAMETERS OF THE --ETC(U)

JUN 82 N J DELANEY, A S PAULSON

DAA629-81-K-0110

UNCLASSIFIED

A-3

ARO-18072.3-MA

NL

1 of 1
SO A
119659



END
DATE
FILMED
11-82
DTIC

ARD 18072.3-MA

12

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER

A-3

2. GOVT ACCESSION NO.

AD-A119 659

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

SELF-CRITICAL, AND ROBUST, ESTIMATES FOR THE
PARAMETERS OF THE MULTIVARIATE NORMAL DISTRI-
BUTION

5. TYPE OF REPORT & PERIOD COVERED

Interim Technical Report

6. PERFORMING ORG. REPORT NUMBER

A-3

7. AUTHOR(s)

N. J. Delaney
A. S. Paulson

8. CONTRACT OR GRANT NUMBER(s)

DAA G29-81-K-0110

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Rensselaer Polytechnic Institute
Troy, New York 1218110. PROGRAM ELEMENT, PROJECT, TASK
AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Approved for public release; distribution
unlimited.

12. REPORT DATE

1 June 1982

13. NUMBER OF PAGES

15

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

Department of the Navy
Office of Naval Research
715 Broadway (5th Floor)
New York, New York 10003

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING
SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

(THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT
ARE THOSE OF THE AUTHOR(S) AND SHOWN AS SUCH. STATED AS
AN OFFICIAL DEPARTMENT OF THE ARMY POSITION OR DE-
CISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

robust estimation, multivariate normal distribution, M-estimators,
outlier identification

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This algorithm yields joint robust estimates of the location vector
and the variance-covariance matrix for samples from the multivariate
normal distribution. The degree of robustness depends on a single filter-
ing parameter, c , set by the user. The algorithm provides, for each
observation, an internally determined weight which may be used to identify
potential outliers.DTIC
SELECTED
SEP 28 1982
H

AD A119659

LINE FILE COPY

SELF-CRITICAL, AND ROBUST, ESTIMATES FOR THE PARAMETERS
OF THE MULTIVARIATE NORMAL DISTRIBUTION

by

N. J. Delaney
Union College

and

A. S. Paulson*
Rensselaer Polytechnic Institute

*Research supported in part by U.S. Army Research Office under contract
DAA G29-81-K-0110.

Self-Critical, and Robust, Estimates for the Parameters
of the Multivariate Normal Distribution

Keywords: Robust estimation; multivariate normal distribution;
M-estimators; outlier identification.

Language: ISO Fortran

Purpose: This algorithm yields joint robust estimates of the location vector and the variance-covariance matrix for samples from the multivariate normal distribution. The degree of robustness depends on a single filtering parameter, c , set by the user. The algorithm provides, for each observation, an internally determined weight which may be used to identify potential outliers.



Accession For	
NTIS	☒
DTIC	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Special
A	

Theory:

The problems inherent in multivariate parameter estimation and the identification of potential outliers have been documented in Gnanadesikan (1977, Chapter 5) and Barnett and Lewis (1979, Chapter 6). Procedures for the simultaneous estimation of the location vector $\underline{\mu}$ and the covariance matrix $\underline{V}$ of the p-variate Gaussian distribution

$$f(\underline{x}|\underline{\mu}, \underline{V}) = \frac{|\underline{V}|^{-1/2}}{(2\pi)^{p/2}} \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{V}^{-1}(\underline{x}-\underline{\mu})\}, \quad (1)$$

given a random sample $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$, have been discussed by Maronna (1976) and Huber (1977, Chapter 5). Our procedure stems from a single underlying primitive principle which reduces to likelihood as a special case and has desirable properties not possessed by competitors. Let c be a real number. The self-critical estimators $\underline{\mu}$ and $\underline{V}$ are determined from the zeros of

$$\sum_{i=1}^n \frac{f_i^c}{Q} \left\{ (1+c) \frac{\partial \log f_i}{\partial \underline{\mu}} - \frac{1}{Q} \frac{\partial Q}{\partial \underline{\mu}} \right\} = \underline{0}, \quad (2)$$

and

$$\sum_{i=1}^n \frac{f_i^c}{Q} \left\{ (1+c) \frac{\partial \log f_i}{\partial \underline{V}} - \frac{1}{Q} \frac{\partial Q}{\partial \underline{V}} \right\} = \underline{0}, \quad (3)$$

where

$$\begin{aligned} Q(\underline{\mu}, \underline{V}; c) &= \int_{R_p} f^{1+c}(\underline{x}|\underline{\mu}, \underline{V}) d\underline{x} \\ &= \{(1+c)^p (2\pi)^{cp} |\underline{V}|^c\}^{-1/2}. \end{aligned} \quad (4)$$

The arguments of f and Q have been suppressed in (2) and (3) for notational convenience. Equations (2) and (3) may also be obtained from the maximization of

$$L_c = \frac{1}{c} \sum_{i=1}^n \left\{ \frac{f_i^c}{Q^{c/(1+c)}} - 1 \right\} \quad (5)$$

with respect to $\bar{\mu}$ and $\bar{V}$ (Paulson and Delaney, 1982). It may be shown that

$$\lim_{c \rightarrow 0} L_c = \sum_{i=1}^n \log f_i \quad (6)$$

so that the self-critical procedure reduces to maximum likelihood as a special case. These procedures may also be developed from consideration of the generalized mean. We have termed the procedure self-critical because the information component supplied by the expressions in brackets $\{\cdot\}$ in (2) and (3) are "fed back" through the assumed density f with degree of criticism determined by c . Observations which receive relatively low final weights f_i^c/Q , a scale-free expression, are candidates for special examination as potential outliers.

Equations (2) and (3) reduce, on simplification, to the joint iterative forms

$$\bar{\mu}_{m+1} = \sum_{i=1}^n w_{mi} x_i, \quad (7)$$

and

$$\bar{V}_{m+1} = (1+c) \sum_{i=1}^n w_{mi} (x_i - \bar{\mu}_m)(x_i - \bar{\mu}_m)^T, \quad (8)$$

where

$$w_{mi} = \frac{\exp\{-\frac{c}{2} (x_i - \bar{\mu}_m)^T \bar{V}_m^{-1} (x_i - \bar{\mu}_m)\}}{\sum_{i=1}^n \exp\{-\frac{c}{2} (x_i - \bar{\mu}_m)^T \bar{V}_m^{-1} (x_i - \bar{\mu}_m)\}} \quad (9)$$

for $m = 0, 1, \dots, m^*$. Initial estimates $\bar{\mu}_0$ and $\bar{V}_0$ are required to set the

iterative procedure embodied in (7), (8), and (9) in motion. We have frequently used w_{m*i} instead of the final f_1^c/Q to assess patterns in the data with respect to the multivariate Gaussian assumption and the internal consistency of the data.

It is recommended that the user experiment with the value c in examining a given data set. The estimators $\bar{\mu}$ and $\bar{V}$ are increasingly robust with increasing c because of the increasingly self-critical nature of the procedure. If for several values of c , similar estimates of $\bar{\mu}$ and $\bar{V}$ are obtained and no weight w_{m*i} is relatively small, then the data and Gaussianity are self-consistent. If the estimates of $\bar{\mu}$ and $\bar{V}$ vary substantially with c or at least one value of w_{m*i} is small relative to the remainder, then those observations so labeled should be set aside for special examination vis-à-vis the remainder. Since the w_{m*i} induce a virtually automatic ranking of the observations, they are very useful in determining patterns in the data. We have typically used $0 \leq c \leq 1$ for the multivariate Gaussian distribution although Paulson, Presser, and Nicklin (1982) have found use for negative values of c , as well as values of c in excess of unity. The magnitude to which c may be set is to a large extent dependent on the sample size.

The estimators $\bar{\mu}$ and $\bar{V}$, for all $c > 0$, are location and scale invariant, are M-estimators, are jointly asymptotic normal, have closed, bounded and redescendent to zero influence functions, and are, for moderate values of c , relatively efficient (Paulson and Delaney, 1982).

Numerical Method:

The algorithm begins by computing initial estimates $\bar{\mu}_0 = \hat{\mu}$, the usual vector of sample means, and $\bar{V}_0 = \hat{V}$, the maximum likelihood estimator of the dispersion matrix. Formulae (7), (8), and (9) are the basis for the self-critical iterative procedure. Successive estimates of μ and V are generated until the norms of consecutive estimates are less than a user specified small amount, ETA, for three successive iterations. Generally, we have used $ETA = 10^{-4}$ or 10^{-5} . Of course, the number of iterations required for convergence will depend on the value of ETA specified.

Structure:

SUBROUTINE SCEST(X,M,N,MDIM,NDIM,C,ETA,IMAX,MU,V,WT,MUZ,VZ,IT,IFALT)

Formal parameters

X	Real array (M,N)	input: data, M dimensions, N points
M	Integer	input: the number of dimensions
N	Integer	input: the number of sample points
MDIM	Integer	input: the row dimension of X in calling program
NDIM	Integer	input: the column dimension of X in calling program
C	Real	input: filtering parameter
ETA	Real	input: convergence criterion for norms of consecutive estimates
IMAX	Real	input: maximum number of iterations desired
MU	Real array (M)	output: final location vector estimate, M dimensions
V	Real array (M,M)	output: final var-cov matrix estimate
WT	Real array (N)	output: final weight assigned to each sample point in computation of estimates
MUZ	Real array (M)	output: initial location estimate; vector of means
VZ	Real array (M,M)	output: initial var-cov estimate; MLE
IT	Integer	output: the number of iterations used
IFALT	Integer	output: 0 if no errors in computation 1 if estimate of var-cov matrix is not positive definite 2 if accuracy test failed on inverse calculation 3 if convergence criteria not met in max desired iterations

Restrictions:

The number of dimensions must not exceed 10, unless the dimensions of dummy arrays VINV, TMP, MUSAV, VSAV, and VNORM are redeclared (see subroutines WEIGHT and CKNRM). This procedure requires a matrix inversion subroutine. As written, subroutine WEIGHT calls on the IMSL (International Mathematical and Statistical Library) subroutine LINV2F which inverts a matrix with specifiable accuracy = IDGT. The use of alternate matrix inversion routines would require rewriting this portion of the subroutine WEIGHT.

Precision:

The DOUBLE PRECISION version is recommended due to the iterative nature of the algorithm and the matrix inverse calculation.

Time:

Experience with an IBM3033 computer has shown that the procedure averaged 0.03 seconds for a bivariate sample of size 20 and 0.08 seconds for a trivariate sample of size 25. These computations were done using DOUBLE PRECISION and a convergence criterion, $\text{ETA} = 10^{-5}$.

References:

1. Barnett, V., and Lewis, T. (1978). Outliers in Statistical Data. New York: Wiley.
2. Gnandesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley.
3. Huber, P.J. (1977). Robust Statistical Procedures. Philadelphia: SIAM.
4. Maronna, R.A. (1976). Robust M-estimators of Multivariate Location and Scatter. Annals of Statistics, 4, pp. 51-67.
5. Paulson, A.S., and Delaney, N.J. (1982). Self-Critical Procedures for the Multivariate Gaussian Distribution. Submitted to Technometrics.
6. Paulson, A.S., Presser, M.A., and Nicklin, E.H. (1982). Self-Critical, and Robust, Procedures for the Estimation of Parameters. Submitted for publication.

A P P E N D I X

FORTTRAN PROGRAM
AND
SUBROUTINES

```

REAL*8 X(5,150),MU(5),V(5,5),WT(150),MUZ(5),VZ(5,5)
REAL*8 ETA,C,CSAV(5),TMP(5),TMP2(5)
DIMENSION VP(30)
NDIM=5
NDIM=150
C DRIVER INPUT:  N  # SAMPLE PTS<= 150,  M  # DIMENSIONS <= 5
C               IMAX MAX ITERS ALLOWED,  NC  # VALUES OF C INPUT
C               ETA CONVERG. CRITERION FOR NORMS
      READ(5,900) N,M,IMAX,NC,ETA
900 FORMAT(4I3,E10.5)
C DRIVER INPUT:  CSAV ARRAY OF C VALUES
      READ(5,901) (CSAV(I),I=1,NC)
901 FORMAT(5F5.3)
C DRIVER INPUT:  VP ARRAY FOR FORMAT OF DATA CARDS PUT IN ()
      READ(5,902) VP
902 FORMAT(30A4)
C DRIVER INPUT:  X(I,J) DATA ONE SAMPLE POINT PER CARD
      DO 10 J=1,M
      READ(5,VP) (X(I,J),I=1,M)
10 CONTINUE
      DO 100 KC=1,NC
      C=CSAV(KC)
      CALL SCEST(X,M,N,NDIM,NDIM,C,ETA,IMAX,MU,V,WT,MUZ,VZ,IT,IFault)
      WRITE(6,910)
910 FORMAT(1H1,17X,20H***** )
      IF(IFault.EQ.0) GO TO 30
C NONNORMAL TERMINATION MESSAGES
      GO TO (11,12,13),IFault
11 WRITE(6,911) IT
911 FORMAT(1H0,46HESTIMATE OF VAR-COV MATRIX ALGORITHMICALLY NOT
      & 19H POSITIVE DEFINITE,4X,10HITER NO. =,I5)
      GO TO 30
12 WRITE(6,912) IT
912 FORMAT(1H0,45HACCURACY TEST FAILED ON INVERSE CALCULATION
      & 10HITER NO. =,I5)
      GO TO 30
13 WRITE(6,913) IMAX
913 FORMAT(1H0,32HCONVERGENCE CRITERIA NOT MET IN ,I5,2X,10HITERATIONS)
C NORMAL TERMINATION AND OUTPUT
30 WRITE(6,930) C,ETA
930 FORMAT(1H0,10X,34H*** SELF-CRITICAL ESTIMATES ***//
      & 24HFILTERING PARAMETER, C =,F7.4,3X,
      & 20HCONVERG. CRITERION =,E12.5)
C COMPUTE INITIAL + FINAL CORRELATION ESTIMATES
C STORE IN LOWER OFF-DIAG OF VAR-COV FOR OUTPUT
      DO 35 I=1,M
      TMP2(I)=DSQRT(VZ(I,I))
      IF(IFault.EQ.0) TMP(I)=DSQRT(V(I,I))
35 CONTINUE
      K=M-1
      DO 45 I=1,K
      L=I+1
      DO 40 J=L,M
      VZ(J,I)=VZ(I,J)/(TMP2(I)*TMP2(J))
      IF(IFault.EQ.0) V(J,I)=V(I,J)/(TMP(I)*TMP(J))
40 CONTINUE
45 CONTINUE
C PRINT OUT INITIAL ESTIMATES

```

```

      WRITE(6,931)
931  FORMAT(1H0,10X,17HINITIAL ESTIMATES)
      DO 50 I=1,M
      WRITE(6,932) MUZ(I), (VZ(I,J),J=1,M)
932  FORMAT(1H0,F10.5,8X,5(F10.5,3X))
      50 CONTINUE
      IF (IT.EQ.1.AND. IFALT.GT.0) GO TO 100
C   PRINT OUT FINAL SC ESTIMATES
      WRITE(6,933) IT
933  FORMAT(1H0,10X,26HFINAL ESTIMATES ITER NO. =,I5)
      DO 60 I=1,M
      WRITE(6,932) MU(I), (V(I,J),J=1,M)
      60 CONTINUE
C   PRINT OUT FINAL SC WEIGHTS FOR EACH SAMPLE POINT
      WRITE(6,934)
934  FORMAT(1H0,3X,3HNO.,3X,6HWEIGHT,10X,12HSAMPLE POINT)
      DO 70 J=1,N
      WRITE(6,935) J,WT(J), (X(I,J),I=1,M)
935  FORMAT(1H0,I5,3X,E12.5,3X,5F10.5)
      70 CONTINUE
      100 CONTINUE
      STOP
      END

```

```

C
C
C      SUBROUTINE SCEST(X,M,N,MDIM,NDIM,C,ETA,IMAX,MU,V,WT,MUZ,VZ,IT,IFAU
C
C      CALCULATES SELF-CRITICAL ESTIMATES
C
C      REAL*8 X(MDIM,NDIM),MU(MDIM),V(MDIM,MDIM),WT(NDIM)
C      REAL*8 MUZ(MDIM),VZ(MDIM,MDIM),ZERO,ONE
C      DATA ZERO,ONE/0.0D0,1.0D0/
C      IFAUIT=0
C
C      OBTAIN INITIAL ESTIMATES- MLE
C      CALL INITL(X,M,N,MUZ,VZ,MDIM,NDIM)
C      DO 10 I=1,M
C      MU(I)=MUZ(I)
C      DO 5 J=1,N
C      5 V(I,J)=VZ(I,J)
C      10 CONTINUE
C
C      ITERATIVE FORMATION OF NEW ESTIMATES
C      DO 100 IT=1,IMAX
C      CALL WEIGHT(X,M,N,MDIM,NDIM,C,MU,V,WT,IFAU
C      IF(IFAU
C
C      LOCATION ESTIMATE
C      DO 20 I=1,M
C      MU(I)=ZERO
C      DO 15 J=1,N
C      15 MU(I)=MU(I)+X(I,J)*WT(J)
C      20 CONTINUE
C
C      VAR-COV ESTIMATE
C      DO 40 I=1,M
C      DO 35 J=1,M
C      V(I,J)=ZERO
C      DO 30 K=1,N
C      30 V(I,J)=V(I,J)+(X(I,K)-MU(I))*(X(J,K)-MU(J))*WT(K)
C      V(I,J)=V(I,J)*(C+ONE)
C      IF(I.NE.J)V(J,I)=V(I,J)
C      35 CONTINUE
C      40 CONTINUE
C
C      CHECK CONVERGENCE CRITERIA
C      CALL CKNPM(M,MDIM,ETA,MU,V,IT,IFLAG)
C      IF(IFLAG.EQ.1) RETURN
C      100 CONTINUE
C
C      CONVERG. CRIT. NOT MET IN MAX ITERATIONS DESIRED
C      IFAUIT=3
C      RETURN
C      END
C
C
C      SUBROUTINE INITL(X,M,N,MUZ,VZ,MDIM,NDIM)
C
C      FORMS INITIAL ESTIMATES- MLE
C
C      REAL*8 X(MDIM,NDIM),MUZ(MDIM),VZ(MDIM,MDIM)
C      REAL*8 DENOM,ZERO,ONE,SUM
C      DATA ZERO,ONE/0.0D0,1.0D0/
C      DENOM=DFLOAT(N)
C      DO 10 I=1,M
C      MUZ(I)=ZERO
C      DO 5 J=1,N

```

```

5 MUZ(I)=MUZ(I)+X(I,J)
  MUZ(I)=MUZ(I)/DENOM
10 CONTINUE
  DO 25 I=1,M
  DO 20 J=1,M
  SUM=ZERO
  DO 15 K=1,N
15 SUM=SUM+X(I,K)*X(J,K)
  VZ(I,J)=(SUM-DENOM*MUZ(I)*MUZ(J))/DENOM
  VZ(J,I)=VZ(I,J)
20 CONTINUE
25 CONTINUE
  RETURN
  END

C
C
C      SUBROUTINE WEIGHT(X,M,N,MDIM,NDIM,C,MU,V,WT,IFAULT)
C
C  FORMS WEIGHTS FOR EACH SAMPLE POINT
C
C      REAL*8 X(MDIM,NDIM),MU(MDIM),V(MDIM,MDIM),WT(NDIM)
C      REAL*8 VINV(10,10),TMP(10,10),WKAREA(50)
C      REAL*8 R,SUM,ZERO,TWO
C      DATA ZERO,TWO/0.0D0,2.0D0/
C      DO 10 I=1,M
C      DO 5 J=1,N
C      5 TMP(I,J)=V(I,J)
C      10 CONTINUE
C
C  OBTAIN INVERSE OF CURRENT VAR-COV MATRIX ESTIMATE
C  LINV2F IS IMSL SUBROUTINE WHICH INVERTS MATRIX WITH
C  SPECIFIABLE ACCURACY=IDGT
C      IDGT=5
C      IER=0
C      CALL LINV2F(TMP,M,10,VINV,IDGT,WKAREA,IER)
C      IF(IER.EQ.0) GO TO 20
C  ERRORS IN MATRIX INVERSION
C      IF(IER.GE.129) IFAULT=1
C      IF(IER.EQ.34) IFAULT=2
C      RETURN
C  END OF SECTION REQUIRED WITH IMSL SUBROUTINE
C
C      20 SUM=ZERO
C      DO 35 J=1,N
C      R=ZERO
C      DO 30 I=1,M
C      DO 25 K=1,N
C      25 R=R+(X(I,J)-MU(I))*VINV(I,K)*(X(K,J)-MU(K))
C      30 CONTINUE
C      R=R*C/TWO
C      IF(R.LT.ZERO) IFAULT=1
C      IF(IFAULT.EQ.1) RETURN
C      IF(R.LT.174.) WT(J)=DEXP(-R)
C      IF(R.GE.174.) WT(J)=ZERO
C      SUM=SUM+WT(J)
C      35 CONTINUE
C      DO 40 J=1,N
C      40 WT(J)=WT(J)/SUM
C      RETURN
C      END

```

```

C
C
SUBROUTINE CKNRM(M,MDIM,ETA,MU,V,IT,IFLAG)
C
C CHECKS CONVERGENCE CRITERIA
C
REAL*8 MU(MDIM),V(MDIM,MDIM),MUSAV(10),VSAV(10,10),VNRM(10)
REAL*8 MUNRM,ZERO
DATA ZERO/0.000/,MCTMU/0/,MCTV/0/
IFLAG=0
IF(IT.EQ.1) GO TO 20
C LOCATION NORM = SUM ABSOLUTE DIFFERENCE OF
C CCNSC. ESTIMATE ELEMENTS
C VAR-COV. NORM = MAX ROW SUM ABS. DIFFERENCE
C CCNSC. ESTIMATE ELEMENTS
MUNRM=ZERO
DO 10 I=1,M
MUNRM=MUNRM+DABS(MUSAV(I)-MU(I))
VNRM(I)=ZERO
DO 5 J=1,M
5 VNRM(I)=VNRM(I)+DABS(VSAV(I,J)-V(I,J))
10 CONTINUE
MAX=1
DO 15 I=1,M
IF(VNRM(I).GT.VNRM(MAX)) MAX=I
15 CONTINUE
IF(VNRM(MAX).LT.ETA) MCTV=MCTV+1
IF(VNRM(MAX).GE.ETA) MCTV=0
IF(MUNRM.LT.ETA) MCTMU=MCTMU+1
IF(MUNRM.GE.ETA) MCTMU=0
C CONVERGENCE CRITERIA MET ON 3 CONSEC. ESTIMATES
IF(MCTV.GT.2.AND.MCTMU.GT.2) IFLAG=1
C SAVE CURRENT ESTIMATES
20 DO 30 I=1,M
MUSAV(I)=MU(I)
DO 25 J=1,M
25 VSAV(I,J)=V(I,J)
30 CONTINUE
RETURN
END

```