DOCUMENT RESUME

ED 425 176	TM 029 238

AUTHOR	Blankmeyer, Eric
TITLE	Robust Means and Covariance Matrices by the Minimum Volume Ellipsoid (MVE).
PUB DATE	1998-05-00
NOTE	7p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Algorithms; Chi Square; *Estimation (Mathematics); *Robustness (Statistics); *Sample Size
IDENTIFIERS	*Covariance Matrices; Mean (Statistics); *Minimum Volume Ellipsoid

ABSTRACT

P. Rousseeuw and A. Leroy (1987) proposed a very robust alternative to classical estimates of mean vectors and covariance matrices, the Minimum Volume Ellipsoid (MVE). This paper describes the MVE technique and presents a BASIC program to implement it. The MVE is a "high breakdown" estimator, one that can cope with samples in which as many as half the observations are contaminated. Samples from a multivariate normal distribution form ellipsoid-shaped "clouds" of data points. The MVE corresponds to the smallest such point cloud containing at least half of the observations, the uncontaminated portion of the data. These "clean" observations are used for preliminary estimates of the mean vector and the covariance matrix. Using these estimates, the program next computes a robust Mahalanobis distance for every observation vector in the sample. Observations for which the robust Mahalanobis distances exceed the 97.5% significance level for the chi-square distribution are flagged as probable outliers. Applications of the MVE are outlined, and a BASIC program is provided so that users can try the algorithm on small or medium data sets before obtaining a more comprehensive version. (SLD)

Robust Means and Covariance Matrices by the Minimum Volume Ellipsoid (MVE)

Eric Blankmeyer
Department of Finance and Economics
Southwest Texas State University
San Marcos, TX 78666
May 1998

Robustness is an issue for any multivariate technique that makes use of mean vectors and covariance matrices. The classical estimates of these statistics are quite vulnerable to outliers. A few bad observations can skew a mean, cause a standard deviation to explode, or distort a correlation coefficient. In their monograph *Robust Regression and Outlier Detection* (Wiley, 1987), P. J. Rousseeuw and A. M. Leroy propose a very robust alternative to the classical estimates --the Minimum Volume Ellipsoid (MVE). This note provides a brief description of the technique and a BASIC program to implement it.

The MVE is a "high-breakdown" estimator; loosely speaking, this means that it can cope with samples in which as many as half the observations are contaminated. (If more than half the data are outliers, no linear estimator can distinguish the good observations from the bad ones !) Samples from a multivariate normal distribution form ellipsoid-shaped "clouds" of data points. The MVE corresponds to the smallest such point cloud containing at least half the observations --the uncontaminated portion of the data. These "clean" observations are used for preliminary estimates of the mean vector (m) and the covariance matrix (S). Using these estimates, the program next computes a robust Mahalanobis distance $(x - m)'S^{-1}(x-m)$ for every observation vector x in the sample. Observations whose robust Mahalanobis distances exceed the 97.5 percent significance level for the chi-square distribution are flagged as probable outliers. (The chi-square statistic has degrees of freedom equal to the number of variables in the sample.) For a detailed discussion of high-breakdown estimation and the resampling algorithm that it uses, please refer to the book by Rousseeuw and Leroy.

The MVE is appropriate for data sets that can reasonably be assumed to come from a multivariate normal distribution (apart from any outliers that may be present). Applications include

- o Hypothesis tests involving means.
- o Hypothesis tests involving covariance or correlation matrices.
- o Linear and quadratic discriminant functions.
- o Identification of high-leverage observations in certain sets of independent variables in logit and probit models.
- o Computation of eigenvalues, eigenvectors, principal components and factor analysis.
- o Canonical correlation.

There are in the public domain several computer implementations of MVE and its variants, including the Minimum Covariance Determinant (MCD) algorithm. Rousseeuw and Leroy's original MINVOL program is available at http://lib.stat.cmu.edu/general/ . The same site contains Douglas Hawkins' "fsa" programs. Rocke and Woodruff's robust estimators are at http://lib.stat.cmu.edu/jasasoftware/ . P. J. Rousseeuw has recently created a new and faster version of MCD which is available at his website: http://win-www.uia.ac.be/u/statis/publicat/fastmcd_abstr. All these programs are FORTRAN or C codes to be compiled by the user. Since most researchers will have ready access to a BASIC interpreter (for example QBASIC or Visual Basic), the attached BASIC version of MVE provides an opportunity to try the algorithm on small or moderate-sized data sets (those having fewer than ten variables) before obtaining a more comprehensive version.

Additional useful references on the MVE and related methods are

P. J. Rousseeuw and B. C. van Zomeren. "Unmasking Multivariate Outliers and Leverage Points," (with discussion) *Journal of the American Statistical Association*, September 1990, Vol. 85, No. 411, 633-651.

D. M. Rocke and D. L. Woodruff. "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, September 1996.

A BASIC PROGRAM TO IMPLEMENT THE MINIMUM VOLUME ELLIPOSID   A1

```
REM  GIVEN A SAMPLE OF N OBSERVATIONS ON K VARIABLES FROM
REM  A MULTIVARIATE NORMAL DISTRIBUTION, SOME OF THE OBSER-
REM  VATIONS MAY BE CONTAMINATED. THIS PROGRAM IDENTIFIES
REM  PROBABLE OUTLIERS IN RELATION TO THE MEAN VECTOR AND
REM  THE COVARIANCE MATRIX. THE USER MAY THEN EXAMINE THE
REM  ANOMALOUS DATA AND ELIMINATE THEM, CORRECT THEM OR
REM  VALIDATE AND RETAIN THEM. THE PROGRAM USES THE
REM  MINIMUM VOLUME ELLIPSOID (MVE) PROPOSED IN THE BOOK
REM  "ROBUST REGRESSION AND OUTLIER DETECTION" BY PETER
REM  J. ROUSSEEUW AND ANNICK M. LEROY (WILEY, 1987).
REM
REM  MVE ESTIMATION IS BASED ON A RESAMPLING SCHEME. THE
REM  MEAN VECTOR AND THE COVARIANCE MATRIX ARE FIT TO A LARGE
REM  NUMBER OF SUBSAMPLES, EACH OF SIZE K+1, AND THE SUBSAMPLE
REM  IS SELECTED WHICH MINIMIZES A CERTAIN FUNCTION OF THE
REM  DETERMINANT OF THE COVARIANCE MATRIX. THIS AMOUNTS TO
REM  CHOOSING THE SMALLEST ELLIPSOID WHICH INCLUDES AT LEAST
REM  HALF THE OBSERVATIONS. THE USER SHOULD SPECIFY
REM  ENOUGH SUBSAMPLES TO PROVIDE VIRTUAL ASSURANCE
REM  THAT THERE WILL BE SEVERAL UNCONTAMINATED SUBSAMPLES.
REM  FOR THIS PURPOSE, THE FOLLOWING GUIDELINES ARE
REM  SUGGESTED:
REM
REM          NUMBER OF                      MINIMUM NUMBER
REM          VARIABLES                      OF SUBSAMPLES
REM              2                              1,000
REM              3                              1,500
REM              4                              2,000
REM              5                              2,500
REM              6 OR MORE                      3,000
REM
REM  FOR FURTHER DISCUSSION, THE USER SHOULD CONSULT THE
REM  BOOK BY ROUSSEEUW AND LEROY.
REM
REM  THE FUNCTION FMED HAS BEEN ADAPTED FROM ROUSSEEUW
REM  AND LEROY'S FORTRAN PROGRAM "PROGRESS".
REM
REM  THERE IS NO WARRANTY, EXPRESSED OR IMPLIED, FOR THIS
REM  PROGRAM. ITS SUITABILITY FOR COMMERCIAL USE OR FOR
REM  ANY PARTICULAR PURPOSE IS NOT GUARANTEED.
REM
3002 DEFLNG I-N
3004 DEFDBL A-H,O-Z
3008 DECLARE FUNCTION FMED (B(), N, NH)
3015 RANDOMIZE TIMER
3020 PRINT "WHAT IS THE INPUT FILE ?"
3025 INPUT INFILE$
3030 PRINT "WHAT IS THE OUTPUT FILE ?"
3035 INPUT OUTFILE$
3040 OPEN INFILE$ FOR INPUT AS #1
3045 OPEN OUTFILE$ FOR OUTPUT AS #2
3050 PRINT "HOW MANY OBSERVATIONS ?"
3055 INPUT N
```

```
3070 PRINT "HOW MANY VARIABLES ?"
3075 INPUT K
3077 K1 = K+1
3080 PRINT "HOW MANY SUBSAMPLES SHOULD BE DRAWN ?"
3085 INPUT ITR
3090 DIM VI(N, K), VJ(K1, K), C(K, K), DIS(N)
3095 DIM AVG(K), M(K1), RDIS(N)
3100 DIM CSMED(10), CS975(10), TEMP(N)
3110 CRITMIN = 1000000#
3115 NH = (N+K1)/2
3120 RK = K
3130 RN = N
3132 RK1 = K1
3135 REM   READ THE DATA
3140 FOR I = 1 TO N
3150 FOR J = 1 TO K
3160 INPUT #1, VI(I, J)
3170 NEXT J
3180 NEXT I
REM READ A TABLE OF THE MEDIAN CHI-SQUARE FOR K = 1-10 D.F.
3250 DATA 0.46,1.39,2.37,3.36,4.35,5,35,6.35,7.34,8.34,9.34
3270 FOR I = 1 TO 10
3280 READ CSMED(I)
3290 NEXT I
REM READ A TABLE OF THE 97.5% CHI-SQUARE FOR K = 1-10 D.F.
3300 DATA 5.02,7.38,9.35,11.14,12.83,14.45,16.01,17.54,19.02,20.48
3310 FOR I = 1 TO 10
3320 READ CS975(I)
3330 NEXT I
REM PERFORM A ROBUST STANDARDIZATION OF THE DATA
3335 N2 = (N+1)/2
3340 FOR J = 1 TO K
3350 FOR I = 1 TO N
3360 TEMP(I) = VI(I,J)
3370 NEXT I
3380 AMED = FMED(TEMP(),N,N2)
3390 FOR I = 1 TO N
4000 VI(I,J) = VI(I,J) - AMED
4010 TEMP(I) = ABS(VI(I,J))
4020 NEXT I
4030 AMED = FMED(TEMP(),N,N2)
4040 FOR I = 1 TO N
4050 VI(I,J) = VI(I,J)/(1.4826*AMED)
4060 NEXT I
4070 NEXT J
REM   START ITERATIONS; CHOOSE A RANDOM SUBSAMPLE OF K1 DATA
4240 FOR L = 1 TO ITR
4250 PRINT "INTERATION  "; L
4260 FOR I = 1 TO K1
4262 M(I) = INT(RND * N) + 1
4264 NEXT I
4266 FOR I = 1 TO K1
4268 FOR J = 1 TO K1
4270 IF I = J GOTO 4274
```

```
4272 IF M(I) = M(J) GOTO 4260
4274 NEXT J
4276 NEXT I
4278 FOR I = 1 TO K1
4280 MI = M(I)
4290 FOR J = 1 TO K
4300 VJ(I, J) = VI(MI, J)
4310 NEXT J
4320 NEXT I
REM   COMPUTE THE SUBSAMPLE COVARIANCE MATRIX
4340 FOR J = 1 TO K
4350 AVG(J) = 0.0
4360 FOR I = 1 TO K1
4370 AVG(J) = AVG(J) + VJ(I,J)
4380 NEXT I
4390 AVG(J) = AVG(J)/RK1
4400 FOR I = 1 TO K1
4410 VJ(I,J) = VJ(I,J) - AVG(J)
4420 NEXT I
4422 NEXT J
4430 FOR I = 1 TO K
4440 FOR J = 1 TO K
4450 C(I,J) = 0.0
4460 FOR JJ = 1 TO K1
4470 C(I,J) = C(I,J) + VJ(JJ,I)*VJ(JJ,J)
4480 NEXT JJ
4485 C(I,J) = C(I,J)/RK
4490 NEXT J
4500 NEXT I
REM INVERT THE SUBSAMPLE COVARIANCE MATRIX
4510 DETC = 1.0
4520 FOR I = 1 TO K
4530 RPIVOT = C(I,I)
4535 IF ABS(RPIVOT) < 0.001 THEN GOTO 4260
4540 DETC = DETC*RPIVOT
4550 C(I,I) = 1.0
4560 FOR JJ = 1 TO K
4570 C(I,JJ) = C(I,JJ)/RPIVOT
4580 NEXT JJ
4590 FOR J = 1 TO K
4600 IF I = J THEN GOTO 4660
4610 CJI = C(J,I)
4620 C(J,I) = 0.0
4630 FOR JJ = 1 TO K
4640 C(J,JJ) = C(J,JJ) - C(I,JJ)*CJI
4650 NEXT JJ
4660 NEXT J
4670 NEXT I
REM COMPUTE THE ROBUST MAHALANOBIS DISTANCES AND FIND
REM THE MEDIAN DISTANCE AND THE MEDIAN ELLIPSOID VOLUME.
4680 FOR I = 1 TO N
4690 DIS(I) = 0.0
4700 FOR II = 1 TO K
4710 FOR JJ = 1 TO K
```

```
4720 DIS(I) = DIS(I)+(VI(I,II)-AVG(II))*(VI(I,JJ)-AVG(JJ))*C(II,JJ)
4730 NEXT JJ
4740 NEXT II
4745 TEMP(I) = DIS(I)
4750 NEXT I
4760 DISMED = FMED(DIS(),N,NH)
4780 CRIT = SQR(DETC*DISMED^RK)
REM IF THIS ELLIPSOID VOLUME IS THE SMALLEST SO FAR, UPDATE
REM THE MEAN VECTOR AND THE COVARIANCE MATRIX.
4790 IF CRIT >= CRITMIN THEN GOTO 4860
4795 CRITMIN = CRIT
4800 FOR I = 1 TO N
4810 RDIS(I) = TEMP(I)
4820 NEXT I
4860 NEXT L
4870 PRINT #2, "97.5% VALUE OF CHI SQUARE = ", CS975(K)
5000 PRINT #2, "OBSERVATIONS WITH ROBUST DISTANCES GREATER"
5010 PRINT #2, "THAN THE 97.5% CHI-SQUARE VALUE (PROBABLE"
5015 PRINT #2, "OUTLIERS)"
5020 FOR I = 1 TO N
5030 RDIS(I) = CSMED(K)*RDIS(I)/DISMED
5040 IF RDIS(I) <= CS975(K) THEN GOTO 5060
5050 PRINT #2, I,RDIS(I)
5060 NEXT I
5070 END

940 FUNCTION FMED (B(), N, NH)
945 DEFLNG I-N
950 DEFDBL  A-H,O-Z
980 LL = 1
990 LR = N
1000 IF LL >= LR GOTO 1210
1010 AX = B(NH)
1020 JNC = LL
1030 J = LR
1040 IF JNC > J GOTO 1180
1050 IF B(JNC) >= AX GOTO 1080
1060 JNC = JNC + 1
1070 GOTO 1050
1080 IF B(J) <= AX GOTO 1110
1090 J = J - 1
1100 GOTO 1080
1110 IF JNC > J GOTO 1170
1120 WA = B(JNC)
1130 B(JNC) = B(J)
1140 B(J) = WA
1150 JNC = JNC + 1
1160 J = J - 1
1170 GOTO 1040
1180 IF J < NH THEN LL = JNC
1190 IF NH < JNC THEN LR = J
1200 GOTO 1000
1210 FMED = B(NH)
1220 END FUNCTION
```

Organization/Address: DEPT FINANCE / ECONOMICS
SOUTHWEST TX STATE UNIV.
SAN MARCOS TX 78666

Telephone: 512-245-3253
FAX: 512-245-3089

E-Mail Address: eb01@business.swt.edu