AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

Comparison between some Parametric Robust Methods for Estimating the Parameters of the Multiple Normal Distribution

M. Murtadha Mansour Abdullah
University of Wasit
College of Administration and Economics
mabdullah@uowasit.edu.iq

Abstract

Estimating the parameters of the Multivariate Normal Distribution is very important process in many statistical Application like the Principal Component Analysis or Canonical Analysis. The paper aims at finding robust and efficient estimators for the parameters of the multivariate normal distribution by using parametric method which is the reweighted minimum vector method RMV and compare it with another robust estimation methods like S estimation method in case of deferent sample sizes and deferent contaminated ratios. Results shows that the reweighted minimum vector is the best method via simulation and real data was taken from waist sewerage directorate minimum esquire error (MMSE) was used as a comparison tool between the two estimation methods.

1. Introduction

Estimating the parameters of the normal distribution is very important and main process **[4]** for many statistical applications and yet it will be more important when we deal with multivariate statistics as the data take deferent features. Estimating the mean vector end the covariance matrix of the multivariate normal distribution is a milestone for many important statistical analysis methods such as factor analysis and discernment analysis the classical methods of estimation location and parameters for the multivariate normal distribution give weak and non-efficient estimates especially when we have large dimension or we have any problem in the assumption of linear regression model to avoid that we use here the RMV and S estimators as new methods for estimating the multivariate Normal Distribution parameters **[7, 11]**

2. The Multivariate Normal Distribution

The multivariate normal distribution function deals with random vectors of multivariate scale units with (n * 1) dimension in a sample of observations let

$\underline{X} = (X_1, X_2 \dots X_n), X_j \epsilon R \ \forall j$ then the multivariate normal distribution function will be **[11]**

$$f(\underline{X}) = \frac{1}{(2\pi)^{\frac{P}{2}} |\Sigma \quad|^{\frac{1}{2}}} \ exp^{-\frac{1}{2}(\underline{X}-\underline{M}) \Sigma X^{-1} (\underline{X}-\underline{M})} \dots \dots .. (1)$$

Where

$\underline{\mu}$: is the mean veaor

Σ : is the var –covariance matrix

3. The Reweighted Minimum Vector Method (RMV)

This method is considered as a parametric method that depends upon giving weights (0) for the extreme values which ($d^2_{RMV} > X^2_{n,0.025}$) and (1) for the no extreme values then the RMV for the scale and location parameters will be **[2,10]**

$$\overline{X}^{ROW}_{RMVV} = \frac{\sum_{i=1}^{M} w_i x_i}{M} \dots \dots \dots \dots .. (2)$$

$$S_{RMVV} = \emptyset^{*\propto}_{m,n,p} \ K^*(M) \frac{\sum_{i=1}^{M} w_i (X_i - \overline{X}^{ROW}_{RMVV})(X_i - \overline{X}^{ROW}_{RMVV})^-}{M} \dots \dots .. (3)$$

Where

d^2_{RMV}: represents the maximum values of the parameters

M : represent the number of observation $d^2_{RMVV} \leq X^2_{n,0.025}$

$$W \begin{cases} 0 & if d^2_{RMVV} \left(X_i - \overline{X}_{rmvv}\right) > X^2_{n,0.025} \\ 1 & o.w \end{cases} \dots \dots \dots \dots .. (4)$$

Which d^2_{RMVV} Will be equal to

$$d^2_{MVV} = (X_i - \overline{X}_{mvv})^{-} S^{-1} (X_i - \overline{X}_{mvv}) \dots \dots \dots (5)$$

Then efficient factor

$$K^*(M) = \frac{m/n}{P(X^2_{P+2} < X^2_{P,\frac{m}{n}})} \dots \dots \dots (6)$$

Last the quadratic distances will be

$$L_i^2(X_i, \overline{X}_{RMVV}, S_{RMVV}) = (X_i - \overline{X}_{rmvv})^{-} S^{-1} (X_i - \overline{X}_{rmvv}) \dots \dots \dots \dots (7)$$

4. S Method

This method considered one of the important methods to have a high robust estimators for the multivariate normal distribution function. Moreover it gives good efficiency prosperity then the S estimator for scale parameter is **[3]**

$$b = \frac{1}{n} \sum_{i=1}^{n} p\{d(X_i, \underline{X}, S)\} \dots \dots \dots \dots \dots \dots \dots (8)$$

Where

$(d(X_i, \overline{Y}, s)$ is the mahalanobis distance which is in form **[1]**

$$d^2 = (X_i - \overline{X})^{-} S^{-1} (X_i - \overline{X}) \dots \dots \dots (9)$$

And for good consistency we have

$$b = E_{F_0}[P(\|X\|)] = E\left(P\left(\frac{d}{k}\right)\right) \dots \dots \dots (10)$$

Where

$F_0 = N(0, I_P)$

And x is cycle random variable

Here we chose Tukey function as weight for pin (8) because it's had a good properties as follow **[3 , 10]**

$$P(MD_{\text{i}}) = \begin{cases} 1-[1-(\frac{MD_{\text{i}}}{k})^2]^3 & if \quad |MD_{\text{i}}| \leq k_0 \\ 1 & if \quad |MD_{\text{i}}| > k_0 \end{cases}$$

Where [MDi] is the Mahalanobis distances and to get high breakdown point we determine the constant (K_o) which gives central property called B-weight estimator and b depends on the value of ko as follow

$$k_0 = \sqrt{p\{\sqrt{\left(\frac{1}{9}\right)\left(\frac{2}{p}\right)c} + (1-\left(\frac{1}{9}\right)\left(\frac{2}{p}\right))\}^3} \dots\dots\dots\dots(11)$$

Then the consistency parameter b can be calculated in the following form in case of normal assumptions **[5, 9]**

$$b = \frac{p}{2}X^2_{P+2}(K_0^2) - \frac{p(p+2)}{2K_0^2}X^2_{P+4}(K_0^2) + \frac{p(p+2)(p+4)}{6K_0^4}X^2_{P+6}(K_0^2) + \frac{k_0^2}{6}\left(1 - X^2_P(K_0^2)\right) \dots\dots\dots\dots(12).$$

5. Minimum Men Square Error (MMSE)

Here we use the MMSE as comparison tool to determine which the best estimator by using the Fibonacci distance as follow **[7,11]**

$$MMSE = E\{\|\varepsilon - \hat{\varepsilon}\|_F^2 = \frac{Trcce(\varepsilon - \hat{\varepsilon})^2}{n} \dots\dots\dots\dots.(13)$$

Then MMSE is good method of comparison among deferent estimators of the var-covariance matrix and for the mean vector. The MMSE for it will simply be the square Euclidean distance in the following from .

$$\|u - r\|^2 = (u_1 - r_1)^2 + (u_2 - r_2)^2 + \cdots + (u_n - r_n)^2 \dots\dots.(14)$$

The MSE for the mean vector will be **[6]**

$$MSE = \frac{1}{P}\left\|\underline{M} - \underline{\widehat{M}}\right\|^2 = \frac{1}{P}((M_1 - \widehat{M}_1)^2 + (M_2 - \widehat{M}_2)^2 + \cdots + (M_n - \widehat{M}_n)^2) \ldots . (15)$$

6. Simulation Results

Box- Muller method were used to generate multivariate normal data that contaminated with deferent ratios with deferent sample sizes by using MIATLAB

We calculate the contaminated distribution with certain (α) the ratio of contaminated and (1- α) is the non- contaminated data as follow **[1 , 8]**

$$X = (1-\alpha)N(M,\sigma^2) + \alpha N(M+t,\sigma^2) \qquad t \neq 0 \quad \ldots\ldots\ldots (16)$$

T is the contaminating degree

We repeat this experiment (1000) time for deferent sample sizes (25 , 70 , 150 , 300) and (α= 0.6 , 0.2 , 0.35) where we get the var – covariance matrix and the mean vector which is the theatrical parameters pop as follow

Then

Table (1) covariance matrix (α = 0.05 , n=25) RMV Method

	X1	X2	X3	X4	X5
X1	**11.52**	**0.68**	**-1.6**	**0.85**	**-4.2**
X2		**25.56**	**-3.38**	**-2.53**	**4.84**
X3			**84.91**	**2.38**	**-4.3**
X4				**110.17**	**7.31**
X5					**242.52**

Table (2) covariance matrix (α= 0.05 , n=25) S Method

	X1	X2	X3	X4	X5
X1	11.41	2.21	-0.55	-3.44	-3.74
X2		62.31	-5.34	-3.51	4.55
X3			74.89	3.65	-4.66
X4				120.11	4.32
X5					232.86

Table (3) covariance matrix (α= 0.05 , n =70) RMV Method

	X1	X2	X3	X4	X5
X1	11.37	0.78	-1.8	-0.3	-3.43
X2		52.2	-3.36	-3.93	5
X3			94.21	4.72	-4.58
X4				100.12	7.21
X5					264.63

Table (4) covariance matrix (α= 0.05 , n=70) S Method

	X1	X2	X3	X4	X5
X1	11.97	0.68	-0.98	-0.72	-4.13
X2		52.4	-3.53	-4.1	-2.77
X3			94.16	2.92	-5.73
X4				110.21	8.7
X5					272.13

Table (5) covariance matrix for contaminated (α=0.05 , n=150) RMV Method

	X1	X2	X3	X4	X5
X1	13.26	0.55	2.7	0.41	4.71
X2		45.88	4.63	4.51	5.91
X3			94.61	4.22	4.88
X4				120.76	8.43
X5					262.21

Table (6) covariance matrix for contaminated (α=0.05 , n=150) S Method

	X1	X2	X3	X4	X5
X1	11.76	0.98	-0.37	-0.96	4.33
X2		62.51	2.75	5.22	4.22
X3			52.61	1.95	3.22

X4				**120.72**	**10.55**
X5					**262.31**

Table (7) covariance matrix for contaminated (α=0.05 , n=300) RMV Method

	X1	X2	X3	X4	X5
X1	**12.36**	**0.86**	**-1.5**	**0.23**	**-4.36**
X2		**45.87**	**5.86**	**-4.35**	**5.98**
X3			**95.31**	**4.17**	**-4.85**
X4				**99.9**	**7.65**
X5					**242.6**

Table (8) covariance matrix for contaminated (α=0.05 , n=300) S Method

	X1	X2	X3	X4	X5
X1	**20.11**	**0.92**	**-0.77**	**-0.63**	**-5.21**
X2		**45.36**	**-5.93**	**-5.93**	**3.12**
X3			**51.97**	**2.55**	**-5.65**
X4				**110.37**	**10.63**
X5					**252.62**

Table (9) mean vector estimator for contaminated normal distribution with (α=0.05 , n=25)

	$\widehat{M}_1$	$\widehat{M}_2$	$\widehat{M}_3$	$\widehat{M}_4$	$\widehat{M}_5$
RMVV	**17.7**	**21.5**	**60.23**	**88.12**	**120.11**
S	**15.24**	**30.29**	**60.11**	**75.31**	**110.11**

Table (10) mean vector estimator for contaminated normal distribution with (α=0.05 , n=70)

	$\widehat{M}_1$	$\widehat{M}_2$	$\widehat{M}_3$	$\widehat{M}_4$	$\widehat{M}_5$
RMVV	**17.22**	**22.15**	**60.75**	**88.41**	**120.04**
S	**15.31**	**30.98**	**60.52**	**76.21**	**110.21**

Table (11) mean vector estimator for contaminated normal distribution with (α=0.05 , n=150)

	$\widehat{M}_1$	$\widehat{M}_2$	$\widehat{M}_3$	$\widehat{M}_4$	$\widehat{M}_5$
RMVV	**17.12**	**26.33**	**60.23**	**85.11**	**118.81**
S	**15.11**	**30.01**	**61.21**	**81.21**	**111.11**

Table (12) mean vector estimator for contaminated normal distribution with (α=0.05 , n=300)

	$\widehat{M}_1$	$\widehat{M}_2$	$\widehat{M}_3$	$\widehat{M}_4$	$\widehat{M}_5$
RMVV	**16.82**	**25.87**	**59.21**	**88.35**	**120.31**
S	**15.26**	**29.91**	**59.81**	**80.21**	**111.21**

Table (13) MSE for covariance matrix for the methods (α = 0.05)

	N	25	70	150	300
%5	**RMVV**	**1.521**	**1.331**	**1.221**	**0.851**
	S	**3.551**	**3.124**	**2.957**	**2.225**

Table (14) MSE for covariance matrix for the methods (α=%10)

	N	25	70	150	300
%10	**RMVV**	**1.211**	**1.031**	**1.001**	**0.671**
	S	**7.521**	**4.231**	**5.211**	**6.321**

Table (15) MSE for covariance matrix for the methods (α=%25)

	N	25	70	150	300
%25	**RMVV**	**3.291**	**5.221**	**2.915**	**0.664**
	S	**12.551**	**28.121**	**11.951**	**9.541**

Table (16) MSE for mean vector for the methods (α= %5)

	N	25	70	150	300
%5	**RMVV**	**1.841**	**1.981**	**1.731**	**1.271**
	S	**3.662**	**4.321**	**4.388**	**2.707**

Table (17) MSE for mean vector for the methods (α=%10)

	N	**25**	**70**	**150**	**300**
%10	**RMVV**	**1.212**	**0.987**	**0.731**	**0.394**
	S	**1.725**	**1.425**	**1.265**	**1.131**

Table (18) MSE for mean vector for the methods (α=%25)

	N	**25**	**70**	**150**	**300**
%25	**RMVV**	**3.491**	**1.521**	**0.845**	**0.562**
	S	**5.211**	**3.744**	**3.211**	**2.451**

From results in the tables above - we can notice that MMSE get smaller with the large sample size for both estimation methods in all contaminated rates the results show that the RMV is the best method to estimate the parameters of the multivariate normal distribution via MSE and its better than S method .

7. Real data

There is many pollution sources for river water and rainwater drainage network is the main sources of pollution. We collect data from the water pollution labrotary in Wasit sewerage directorate in Wasit state by taking samples of water from deferent spots in 51 week by helping hand from the employees of Wasit sewerage directorate we determine 5 variables as a pollution sources.

X_1: O8G represent the oil and greases in water

X_2: SO4 represent the sulfates in water

X_3: AIK represent the alkaline in water

X_4: PH represent the acidic in water

X_5: TOS represent the total dissolved salts in water

Where the data tested to normal distribution and we estimate the location and the scale parameters for the multivariate normal distribution by using RMV and S methods by using MATLAB program and we calculate the MSE for covariance matrices and the mean vectors as follow

Table (19) estimated scale parameters by using RMV

	X1	X2	X3	X4	X5
X1	100.25	-1.55	1.62	62.89	44.55
X2		210.51	31.55	-59.25	-110.21
X3			562.11	-259.51	151.22
X4				975.24	-582.11
X5					1251.22

Table (20) estimated scale parameters by using S method

	X1	X2	X3	X4	X5
X1	100.15	-1.76	1.5	64.21	40.22
X2		210.93	36.22	-63.96	-113.21
X3			562.21	-267.11	149.22
X4				975.19	-552.11
X5					1251.10

Table (21) estimated mean vector by using RMV and S method

	$\widehat{M}_1$	$\widehat{M}_2$	$\widehat{M}_3$	$\widehat{M}_4$	$\widehat{M}_5$
RMVV	**8.51**	**35.51**	**60.91**	**100.11**	**40.21**
S	**5.11**	**42.48**	**80.22**	**110.22**	**180.22**

Table (22) MSE for estimated scale prameters by using RMV and S method

RMVV	S
1.229	**7.521**

Table (23) MSE for estimated mean vector by using RMV and S method

RMVV	S
0.421	**2.551**

8. Conclusion

We can see that the value of MSE for covariance matrix is decreasing as the sample size get larger and the results show that the RMV method is better than S method to estimate the covariance and the mean vector in small and large sample size we recommend the method of RMV as an efficient method to estimate location and scale parameters for the multivariate normal distribution

9. References

1- Ahmed, A. B. (2008). Standard Modeling of National Energy Consumption in Algeria during the Period (1988: 10-2007: 03). (Master Unpublished). University of Algeria, Faculty of Economic Sciences and Management Sciences.

2- Olive, D. J. (2018) .Robust Multivariate analysis .springer.

3-Ali, H. Yahiya, s. s. s. & Omar, Z. (2014, June). The efficiency of reweighted minimum vector variance. In AIP conference proceedings (Vol. 1602, NO. 1, pp.1151-1156).AIP.

4- Bable, B., & Pawar, D. (2012). Vector time series: the concept and properties to the vector staionary time series. International Research Journal of Agricultural Economics and Statistics, 3(1), 84-95.

5- Verardi, V. & Mc Cathie, A. (2012). The S-estimator of multivariate Location and Scatter in Stata. Stata journal, 12(2), 299

6- Huang, S.-C. (2008). Combining wavelet-based feature extractions with SVMs for financial time series forecasting. Journal of Statistics and Management Systems, 11(1), 37-48.

7- Hossjer, O., Croux, C, & Rousseeuw, p. j. (1994) Asymptotic of generalized s-estimation journal of Multivariate Analysis, 51(1), 148-177.

8-Croux, C. Rousseeuw, p. J & Hossjer, O. (1994). "Generalized s-estimator", JASA, Vol. 89, NO. 428, 1271-1281.

9-Bigot, J, Biscay, R. J.loubes, J. M. & Muniz-Alvarez, L. (2011) .Group lasso estimation of high-dimensional covariance matrices. Journal of Machine learning Research, 12(Nov), 3187-3225

10-Chen, Y. , Wiesel, A. Eldar, Y. C. , & Hero, A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. IEEE Transactions on signal processing, 58

(10),5016-5029

11- Iman, T. M. (2014). Standard analytical study of family consumption of electricity - Study of the case of Sonelgaz Unit Al Buira - during the period 2008:1 - 2013:12 (Vol. 18). Ministry of

Higher Education and Scientific Research University of Akli Mahnood Olhad/Faculty of Economic and Commercial Sciences and Management Sciences.