JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-19, No.-5, May (2024) pp 73-80 ISSN (Print) 0973-8975

COMPARISON MRCD AND ORACLE FOR ESTIMATING THE DETERMINANT OF HIGH DIMENSIONAL COVARIANCE MATRIX

Fatimah Abdul – Hammeed Jawad Al – Bermani[1], Mohammad Huseen Abdul – Hammeed Jawad Al – Bermani[2]

[1]Department of Statistics, Administration and Economy College, University of Baghdad, Iraq.

[2]Department of Basic Sciences, Agricultural College, University of Baghdad, Iraq.

Email: [1]Fatimah.a@coadec.uobaghdad.edu.iq
[2]Mohammed.h@coagri.uobaghdad.edu.iq

Corresponding author: **Fatimah Abdul–Hammeed Jawad Al–Bermani**

https://doi.org/10.26782/jmcms.2024.05.00005

(Received: February 24, 2024; Revised: April 11, 2024; Accepted: April 26, 2024)

Abstract

Estimating the variance matrix has an important role in statistical applications and conclusions, in high–dimensional matrices if the number of variables is greater than the number of observations P > n, the traditional statistical methods are not reliable because they give uncontrolled estimates. Shrinkage methods are used to estimate the high–dimensional variance matrix.

In this research, the high–dimensional variance matrix was estimated using the robust Nonparametric method Minimum Regularized Covariance Determinant (MRCD), which is based on Mahalanobis distance, and compared with the variance matrix estimated by the Oracle method, which is based on the Frobenius criterion.

Keywords: Frobenius, High–Dimensional, Minimum Regularized Covariance Determinant, Mahalanobis, Oracle, Parameter regulation, Shrinkage,

I. Introduction

There are some cases that the researcher faces in estimating the variance matrix when the number of variables (p) is greater than the number of observations (n), they are called high–dimensional data it has become common especially when the study is about cancer, clinical, financial and signal processing … etc.

When P>n, the usual methods are inefficient in estimating the variance matrix, so the shrinkage method is used in estimating the variance matrix,S=δT+(1-δ)S where T is the target matrix and δ is the shrinkage coefficient 0$\leq \delta \leq$ 1 [XIX].

To find the best estimation of the high–dimensional variance matrix, the systematic variance estimation (MRCD) was developed for estimating the variance matrix method, which is a modification of the method (MCD), which is used for multivariate and robust Nonparametric data.

By finding the subset h using the smallest distance computed from the regular covariance to estimate the variance matrix.

S = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)(x_i - \bar{x}\,)'$, In the first section, a method was used to estimate the variance matrix when p>n it is a generalization to (MCD) it is called the method of estimating the regularized covariance (MRCD), and it is computed from Mahalanobis distance using the regular covariance and finding the subset (h).In section 2, the variance matrix was estimated in the high–dimensional data using the shrinkage method, depending on the Frobenius distance. $\hat{\Sigma}$ = (1-u)s +uF, $\hat{\Sigma}$ shrinkage covariance ,u shrinkage intensity 0<u<1 ,F shrinkage Target, F = $\frac{tr(\Sigma)}{p}$,p representing dimension. the idea of the shrunken it to reduce the risk function E{$\left\|\hat{\Sigma} - \Sigma\right\|^2$ } =tr$(\hat{\Sigma} - \Sigma)^2$ [XVII].

II. Minimum Estimator of Regularized Covariance Determinant (MRCD)

Estimation depends on MRCD on the target matrix T and parameter regulation.

The regularity covariance matrix is estimated for a set h of sub-data using the smallest computed distance of regular covariance and an algorithm c-step, is applied where the estimation MRCD depends on the target T matrix and regularization parameter.

$$T=I_p \tag{1}$$

I represents the identity matrix with p variables and [XIX], T=R_c , that is

$$R_c = cj_p+(1-c)I_p \tag{2}$$

j ones matrix , -1/(p-1) < c < 1

As for the regular covariance matrix, which depends on h [XIV].

$$K_{RCM}(H) = \rho T + (1-\rho)\, c_\alpha S_u(H) \quad , 0< \rho < 1 \tag{3}$$

Where $S_u(H) = \frac{(x_H - \mu_x)'\,(x_H - \mu_x)}{h-1}$, c_α consistency factor , 0$< \rho <$ 1 , h = $\frac{3n}{4}$

$$\mu_1 = \frac{1}{h} \sum_{i \in H_1} x_i \quad (4)$$

$$s_1 = \frac{1}{h} (x_i - \mu_1)(x_i - \mu_1)' \quad (5)$$

$$k_{RCM}\,(1) = \rho \mathrm{T} + (1\text{-}\rho) s_1 \quad (6)$$

$$d_{RCM}\,1(i) = (x_i - \mu_1)'\, k_{RCM}^{-1}\,(1)\,(x_i - \mu_1) \quad \text{i=1,2,…,n} \quad (7)$$

Where $d_{RCM}\,1(i)$ Mahalanobis distance

$$\sum_{i \in H_2} d_{RCM}\,1(i) \leq \sum_{i \in H_1} d_{RCM}\,1(i)$$

$$\mu_2 = \frac{1}{h} \sum_{i \in H_2} x_i \quad (8)$$

$$s_2 = \frac{1}{h} (x_i - \mu_2)(x_i - \mu_2)' \quad (9)$$

$$k_{RCM}\,(2) = \rho \mathrm{T} + (1\text{-}\rho) s_2 \quad (10)$$

det $(k_{RCM}\,(2)) \leq$ det $(k_{RCM}\,(1))$, $\mu_2 = \mu_1$, $k_{RCM}\,(2) = k_{RCM}\,(1)$

To find the Minimum Regularized Covariance Determinant (MRCD) by using the c-step algorithm as follows:

- estimate's m_i and scatter estimates S_i, to obtain 6 robust were (i=1,2,…,6).

- determine H_{0i} containing h observation for minimum Mahalanobis distance d_i.

- determine the smallest value of $(0 \leq \rho_i < 1$ where $\rho_i I$+ (1-ρ_i)$c_\alpha S_w(H_{0i})$, if $max_i\{0.1; median_i \rho_{0i}\}$ This leads to 6 different regularization parameters ρ.

- The initial subset H_{0i} for wich $\rho_{0i} \leq \rho$ repeat c-steps from using ρI + (1-ρ) $c_\alpha S_w$ (H_{0i}) until convergen.

- the lowest determinant of ρI + (1-ρ)$c_\alpha S_w(H_i)$ be the subset.

- compute MRCD location and scatter estimates.

II. The Oracle estimator

Shrinkage methods are often used to estimate the variance matrix in applications where the number of variables is greater than the number of observations p>n. Oracle estimate was used to estimate the high–dimensional covariance matrix based on Frobenius distance.

The Oracle estimate $\hat{\Sigma}_0$:

$$\min_{\rho} E\{ \left\| \hat{\Sigma}_0 - \Sigma \right\|_F^2 \} \quad (11)\ \text{[III]}$$

$$\mathrm{S}\ .\ \mathrm{t}\ \hat{\Sigma}_0 = (1\text{-}\rho_0)\mathrm{s} + \rho_0 \mathrm{F} \quad (12)$$

$$\rho_0 = \frac{E\{Tr(\Sigma-\hat{s})(\hat{F}-\hat{s})\}}{E\{\|\hat{s}-\hat{F}\|_F^2\}}$$

$$\rho_0 = \frac{\left(1-\frac{2}{p}\right)Tr(\Sigma^2)+Tr^2(\Sigma)}{\left(n+1-\frac{2}{p}\right)Tr(\Sigma^2)+\left(1-\frac{n}{p}\right)Tr^2(\Sigma)} \quad (13)$$

Is specifies the optimal shrinkage coefficient

$$E\{Tr((\Sigma-\hat{s})(\hat{F}-\hat{s}))\} = \frac{Tr(\Sigma)}{p} E\{Tr(\hat{s})\} - \frac{E\{Tr^2(\hat{s})\}}{p} - E\{Tr(\Sigma\hat{s})\} + E\{Tr(\hat{s}^2)\}$$

$$E\{\left\|\hat{s}-\hat{F}\right\|_F^2\} = E\{Tr(\hat{s}^2)\} - 2E\{Tr(\hat{s}\hat{F})\} + \{Tr(\hat{F}^2)\}$$

$$= E\{Tr(\hat{s}^2)\} - \frac{E\{Tr^2(\hat{s})\}}{p}$$

$$E\{Tr(\hat{s})\} = Tr(\Sigma) \quad (14)$$

$$E\{Tr(\hat{s}^2)\} = \frac{n+1}{n} Tr(\Sigma^2) + \frac{1}{n} Tr^2(\Sigma) \quad (15)$$

$$E\{Tr^2(\hat{s})\} = Tr^2(\Sigma) + \frac{2}{n} Tr(\Sigma^2) \quad (16)\ [XVII]$$

III . Simulation

Data were generated for sample sizes (125*125), (125*150),(150*250) to estimate the determinant of high–dimensional matrix was applied using the target function T=I_p, T=R_c and regularization parameter ρ=0.5 by using the Regularized Minimum covariance Determinant and Oracle Estimator for Frobenius distance.

Table 1: the determinant of variance-covariance matrix by using Oracle Estimator and MRCD

	Determinant		
	125x125	125x150	150x250
Oracle	2.3087e-128	-4.9306e-19	-3.4821e-20
MRCD T=I_p	6.3640e-11	4.8994e-13	2.2263e-21
MRCD T=R_c	5.9477e-42	9.7745e-51	9.1847e-86

IV . Discussion

The study showed the efficiency of the Oracle Estimator by comparing covariance determinants when n=p, n<p.

When applying the genetic Algorithm MRCD, showed efficiency for using T=R_c than T=I_n.

As for the real data, the experiment was conducted on the plot of land n=2, and variables p=14, which were represented by organic fertilizers added to the soil and observing the effect of the organic fertilizers on the content of the level of the tomato crop of Nitrogen (N) , Phosphorous (P) , and Potassium (K) .

Table 2 : real data for Oracle estimator and MRCD

	Variance			
		MRCD T=I_p	MRCD T=R_c	Oracle
Cow manure 5%	N	0.7525	0.5049	0.0244
	P	0.7503	0.5006	0.0010
	K	0.7532	0.5064	0.0165
Cow manure+humic acid spray	N	0.7536	0.5072	0.0267
	P	0.7500	0.5000	4.4643e-04
	K	0.7566	0.5132	0.0233
Cow manure+ adding humic acid	N	0.7528	0.5056	0.0251
	P	0.7505	0.5009	0.0013
	K	0.7598	0.5196	0.0297
Cow manure+spraying and adding humic acid	N	0.7731	0.5462	0.0657
	P	0.7502	0.5004	8.2143e-04
	K	0.7572	0.5144	0.0245
sheep manure 5%	N	0.7541	0.5081	0.0276
	P	0.7502	0.5004	8.2143e-04
	K	0.7502	0.5004	0.0105
sheep manure+humic acid spray	N	0.7566	0.5132	0.0327
	P	0.7501	0.5002	6.4643e-04
	K	0.7598	0.5196	0.0297
sheep manure+adding humic acid	N	0.7561	0.5121	0.0316
	P	0.7502	0.5004	8.2143e-04
	K	0.7528	0.5056	0.0157
sheep manure+spraying and adding humic acid	N	0.7506	0.5012	0.0207
	P	0.7503	0.5006	0.0010
	K	0.7501	0.5001	0.0102
Poulty manure 5%	N	0.7578	0.5156	0.0351
	P	0.7501	0.5002	6.4643e-04
	K	0.7605	0.5210	0.0311
poulty manure+humic acid spray	N	0.7671	0.5342	0.0537
	P	0.7500	0.5000	4.2143e-04
	K	0.7545	0.5090	0.0191
poulty manure+adding humic acid	N	0.7950	0.5900	0.1095
	P	0.7500	0.5000	4.4643e-04
	K	0.7566	0.5132	0.0233
poulty manure+spraying and adding humic acid	N	0.7506	0.5012	0.0207
	P	0.7501	0.5002	6.4643e-04
	K	0.7506	0.5012	0.0113
Chemical fertilization	N	0.7636	0.5272	0.0467
	P	0.7501	0.5002	6.4643e-04
	K	0.7528	0.5056	0.0157
Without fertilizing	N	0.7528	0.5056	0.0251
	P	0.7508	0.5016	0.0020
	K	0.7561	0.5121	0.0222
Determinant	N	0.0211	7.4742e-08	4.2543e-20
	P	0.0179	5.7018e-08	4.4561e-20
	K	0.0195	6.4054e-08	4.2545e-21

The results showed by comparing the variances and the determinant of the covariance matrix the efficiency of Oracle estimating the high–dimensional data, as for to comparison between MRCD for T=I_p, T=R_c that used regulation parameter ρ=0.75 , c=0.5.

The determinant of MRCD T = R_c less than the determinant of MRCD T=I_p , and the

R_c =cj_p + (1-c)I_p , if c=0 then R_c = I_p

The experiment was conducted on two plots of land to which organic and chemical fertilizers were added and their effect on the leaves of the tomato crop of Nitrogen, Phosphorous and Potassium .

As for the variation in the content of leaves when adding organic and chemical fertilizers, there is a small effect in containing Nitrogen, Phosphorous, and Potassium when applying MRCD, as for Oracles' estimate, there was a small effect on the leaves' containment of Phosphorous through the variations, as well as the lowest determinant of variance matrix and for all methods in the leaves' containment of Phosphorous.

V . Conclusion

Using the Oracle method based on a Frobenius distance gave less variation than other methods.

Fertilization with organic and chemical materials gives a high percentage of leaves containing phosphorus. A comparison was also made between fertilization with organic and chemical materials and not adding fertilizer to the soil, as it reduces the percentage of phosphorus in the leaves as well as the percentage of potassium and nitrogen.

Conflict of Interest:

The author declares that there was no conflict of interest regarding this paper.

References

I. F. Abdul–Hammeed, and M. Sabah, : 'Compared with genetic algorithm Fast-MCD-Nested Extension and neural network multilayer Back propagation'. *JOURNAL OF ECONOMIC & ADMINISTRATIVE SCIENCE*. Jun. No 22(89), 381-395, (2016).

II. F. Virgile, V. Gael, T. Benjamin, and Bertrand Thirion. : 'Detecting outlying Subjects in High-Dimensional Neuroimaging Datasets with Regularized Minimum Covariance Determinant'. pp. 264-271. https://hal.inria.fr/inria-00626857. 10.1007/978-3-642-23626-6_33

III. I. Clifferd, : 'High Dimensional Covariance Matrix Estimation'. Department of Statistics, London School. http://stats.lse.ac.uk.

IV. J. Brian Williamson, : 'Shrinkage Estimators for high-dimensional Covariance matrices'. POMONA COLLEGE , April 4, (2014). 10.1109/ICASSP.2009.4960239

V. K. Jan, and H. Jaroslav, : 'Robust Regularized Discriminant Analysis Based on Implicit Weighting'. Technical report No.v-1241. December (2016). http://www.nusl.cz/ntk/nusl-262425

VI. K. Jan, T. Jurjen Duintjer, and S. Anna, : 'Robustness of High-Dimensional Data'. Mining.Kalina@cs.cas.cz. https://www.semantis/scholarory

VII. L. Olivier, W. Michael, : 'Shrinkage Estimation of large covariance matrices: keep it simple'. statistician. university of Zurich . *Journal of Multivariate Analysis*. 186, (2021) 104796. 10.1016/j.jmva.2021.104796

VIII. M. Abdul – Hammeed,and F. Abdul – Hammeed, : 'Estimated between the two-stage summation shrinkage for the variance of a normal distribution and for equal sizes of the two samples'. *Baghdad science journal*. Jun. No 1009, (2011).

IX. M. Hubert, and M. Debruyne, : 'Minimum Covariance Determinant'. *Wiley Interdisciplinary Reviews:Computional Statistics*. 2(2010). Pp.-36-34. https://wis.kuleuven.be/stat/robust/papers/2010/wire-mcd.pdf

X. O. Ledoit ,and M. Wolf , : 'Quadratic Shrinkage for Large Covariance Matrices'. University of Zurich , November (2019). http://dx.doi.org/10.2139/ssrn.3486378

XI. O. Ledoit ,and M. Wolf. : 'A well-conditioned estimator for large-dimensional covariance matrices'. *Journal of Multivariate Analysis*. 88(2) (2004), pp. 365-411. 10.1016/S0047-259X(03)00096-4

XII. R. Maronnan, and R.H. Zamar. : 'Robust Estimates of Location and Dispersion for High-Dimensional Datasets'. *Technometrics*. 44(4), 307-317 (2002). https://www.jstor.org/stable/1271538

XIII. P. Rousseeuw , S . Vanduffel and T. Verdonckl. : 'Minimum Regularized Covariance Determinant Estimatimater'. june 1. (2018). **

XIV. P. Rousseeuw, V. Steven, and V. Tim. : 'The Minimum Regularized Covariance Determinant Estimator'. ar Xiv:1701.07086v3, November 29 (2018). 10.2139/ssrn.2905259

XV. P. Rousseeuw, and D. Van. : 'Afast algorithm for the Minimum Covariance Determinant estimator'. *Technometrics*. 41(3), (1991), pp. 212-223. doi.org/10.2307/1270566

XVI. Won J. H, Lim J. Kim S., J. Rajaratnan. : 'Condition-number regularized covariance estimation'. *J. R. Stat. Ser B (stat.Methodol)* 75 (3), (2013) 427-450. doi.org/10.1111/j.1467-9868.2012.01049.x

XVII. Yilun C., Ami wiesel, Alfred O. Hero III. : 'Shrinkage Estimtion of high Dimensional Covariance Matrices'. *International Conference on Acoustics, Speech and Signal Processing*. April (2009) 10.1109/ICASSP.2009.4960239

XVIII. Yilun C., Ami Wiesel, Alfredo. : 'Robust Shrinkage Estimtion of high Dimensional Covariance Matrices'. arXiv:1009.5331v1 [stat.ME]. 27 sep (2010). 10.1109/TSP.2011.2138698

XIX . Zongliang Hu, Kai Dong, Wenlin Dai and Tiejan Tong. : 'Acomparison of Methods for Estimating the Determinent of High-Dimensional Covariance Matrix'. *The International Journal of Biostatistics.* September, (2017). doi.org/10.1515/ijb-2017-0013