https://doi.org/10.31272/jae.i147.1309
https://admics.uomustansiriyah.edu.iq/index.php/admeco
P-ISSN: 1813–6729 E-ISSN: 2707-1359

High-Dimensional Variance Matrix Estimation Using the OGK Genetic Algorithm

Fatimah Abdul – Hammeed Jawad AL-Bermani
Dep. of Statistics, Administration &Economy College, University of Baghdad, Iraq, Baghdad.
Email: Fatimah.a@coadec.uobaghdad.edu.iq, ORCID ID:/ https://orcid.org/0000-0001-8767-1410

Article Information

Article History:
Received: 26 / 11 / 2024
Accepted: 9 / 1 / 2025
Available Online: 1 / 3 / 2025
Page no: 24 – 29

Keywords:
regularization parameter, minimum determinantofregularization variance, Target Matrix Mahalanobis distance, High-dimensional

Correspondence:
Researcher name:
Fatimah Abdul – Hammeed jawad

Email:
Fatimah.a@coadec.uobaghdad.edu.iq

Abstract

This research estimated a high-dimensional variance matrix when the number of variables was more significant than the number of observations. The OGK genetic algorithm was applied to find the variance matrix. A modification for the genetic algorithm OGK was proposed depending on the regular parameter ρ and the target matrix T, and It was called the (Orthogonalized Regularized Gnanadesikan – Kettenring) it can be written briefly (ORGK), The data was taken from four stations representing the monthly rates for a group of polluted for air the gases for one year. The monthly rates were measured for four types of gases (Methane gas CH_4 , Carbon Monoxide gas CO , Nitrous gas NO_2 , Sulfur Dioxide gas SO_2) . In this study, a comparison was made between the genetic algorithm OGK, ORGK and MRCD by finding the determinant of the covariance matrix and identifying the most polluting gasses. Carbon Monoxide CO was the main cause of pollution. And the ORGK algorithm dependent of the regular parameter and target matrix, it has a clear effect in obtaining the lowest determinant of the covariance matrix, which is called (Orthogonalized Regularized Gnanadesikan – Kettenring) ORGK

1. Introduction

In many studies, the variance matrix is often used to compare between methods or to show the most influential factors. In higher dimensional data when p>n is evident in studies related to cancer research, financial mathematics, signal and image processing, criminology, and clinical trials [6]. Where the genetic algorithm (Minimum Regularized Covariance Determinant) MRCD was used to estimate the minimum determinant of the covariance matrix, which depends on the parameter regulation and is denoted by ρ and the target matrix T (Target Matrix), which gives the best estimate from the usual methods and is also considered a generalization of the MCD algorithm[10].

$$X = (x_1, x_2, \dots., x_n)' \quad (1)$$

X: matrix n*p

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})' \quad (2)$$

$$S = \rho T + (1-\rho)s_1 \quad (3)$$

If $\rho = 0$, then the covariance matrix using the regularisation parameter and the target matrix is equal to the covariance matrix, that is, $S = S_1$ [7] .

In 2011, researchers (Mia Hery, Peter J. Rousseeuw, and Tim Verdonck) introduced an algorithm that relied on the correlation matrix to estimate robust covariance, find eigenvectors, and compute the positive definite matrix [10].

In the same year, researchers (Virgile Fritsch, Gaël Varoquaux, Thyreau Benjamin, Jean-Baptiste Poline, and Bertrand Thirion) incorporated a regularization parameter to find the minimum determinant of covariance. This approach provided better estimates due to numerous outliers and high-dimensional data. The method was applied to medical imaging in clinical studies and brain imaging, where the data was characterised by high variability and dimensionality [12].

In 2018, researchers (Peter J. Rousseeuw, Steven Vanduffel, and Tim Verdonck) introduced the target matrix TTT and the regularization parameter ρ\rhoρ. They employed regularised covariance to compute Mahalanobis distances in the MRCD method, resulting in simpler solutions and more accurate results than the MCD method [9].

In this study, the OGK algorithm was further developed. The researcher focused on estimating the covariance matrix using the regularisation parameter and the target matrix, naming the new approach the ORGK algorithm.

The study aims to apply genetic algorithms to identify the most significant key factors influencing air pollution and to compare methods by determining the determinant of the covariance matrix.

2. High-dimensional variance matrix estimation using genetic algorithms:

2.1 MRCD (Minimum Regularized Covariance Determinant) Genetic Algorithm

The MRCD algorithm is a generalization to the MCD algorithm where it is applied when p>n to get more accurate estimates that depend on the regulation parameter ρ(parameter regulation), the target matrix T (target matrix), and the Mahalanobis distance. Here are the steps of the genetic algorithm MRCD [14]:

1- The data is divided into h=3n /4 to obtain the minimum distance.

2- We find the estimate of the mean m_i and the variance S_i , i= 1,2,…,6 to obtain estimates as :

$$S_1 = \frac{1}{h}\sum_i^h (x_i - m_1)(x_i - m_1)' \quad (4)$$

$$m_1 = \frac{1}{h}\sum_i^h x_i \quad (5)$$

3- We find the positive defined matrix

$$S_R(1) = \rho \mathrm{T} + (1-\rho)\, c_\alpha S_1 \quad (6)$$

that is 0 <ρ <1 , -1/(p-1) < c <1 , T=I_p

4-

$$d_R\, 1(i) = \sum_i^h (x_i - m_1)'\, S_R^{-1}(1)\, (x_i - m_1) \quad (7)$$

Where $d_R\, 1(i)$ Mahalanobis distance

5-

$$S_2 = \frac{1}{h_2}\sum_i^h (x_i - m_2)(x_i - m_2)'\, m_2 = \frac{1}{h_2}\sum_i^h x_i \quad (8)$$

6- We find the positive defined matrix

$$S_R(2) = \rho \mathrm{T} + (1\,\rho) c_\alpha S_2 \quad (9)$$

7-

$$d_R\, 2(i) = \sum_i^h (x_i - m_2)'\, S_R^{-1}(2)\, (x_i - m_2) \quad (10)$$

8-

$$\sum_{i \in h_2} d_R\, 2(i) \le \sum_{i \in h_1} d_R\, 1(i) \quad (11)$$

9- The steps are repeated until the robust estimates of 6 are obtained and the lowest distance, which represents the S_R variance matrix, is obtained [12].

2.2 OGK (Orthogonalized Gnanadesikan – Kettenring) Genetic Algorithm

In robust estimates, the genetic algorithm is applied to find the variance matrix when the number of observations exceeds the number of variables n>p, which gives a robust estimate. Here are the steps of the OGK genetic algorithm [4]:

1- We find an estimate of the mean m_i, and the variance, S_i

2- Find $y_i = D^{-1}x_i$ where i=1,2,…,n

D=diag(S(x_1), S(x_2),…, S(x_p)) (12)

3- We find the correlation matrix U where

$u_{JK} = \frac{1}{4}\,(\mathrm{S}(Y_J + Y_K)^2 - \mathrm{S}(Y_J - Y_K)^2)$ (13)

$\mathrm{Y} = (Y_1, Y_2, \dots, Y_p)$

4- We find the matrix E of the characteristic vector to U and we find the following:

A) We find the characteristic vector $V = YE$

B) We find the robust variance $V = (V_1, V_2, \dots, V_p)$

$\Lambda = \mathrm{diag}\,(s^2(V_1), \dots, s^2(V_p))$ (14)

C) We find $\mu(\mathrm{Y}) = \mathrm{E(m)}$ where $\mathrm{m} = (m(V_1), \dots, m(V_p))'$, $\hat{\Sigma}(\mathrm{Y}) = \mathrm{E}\Lambda E'$

So mean estimate $\hat{\mu}_{OGK}(\mathrm{x}) = \mathrm{D}\hat{\mu}(\mathrm{Y})$ and variance

$\hat{\Sigma}_{OGK}(\mathrm{x}) = \mathrm{D}\hat{\Sigma}(\mathrm{Y})D'$

Describe FastMCD and OGK algorithms to estimate the covariance matrix, use aspects of both, and call it DetMCD [9].

But when p>n, that is, the number of variables is greater than the number of observations, which are called high-dimensional data, then the estimation of the high-dimensional variance matrix using the usual methods gives inaccurate estimates, so the researcher proposed a new algorithm based on the regulation parameter ρ,

And the target function T is called ORGK, where the MRCD algorithm was used as an initial for estimate, then the steps of the OGK algorithm are followed, which are as follows:

1- The mean m_i and the variance S_i are estimated as a preliminary estimate, then the variance matrix is estimated as follows:

$S_R = \rho\mathrm{T} + (1\rho) - c_\alpha S_i, \quad 0 < \rho < 1\ , \quad -1/(\mathrm{p}-1) < \mathrm{c} < 1 < \mathrm{c} < 1\ , I_p$ (15)

2- We find $y_i = D_R^{-1}x_i$ where $i = 1,2, \dots, n$

$D\_R = diag(s\_R\ (x\_1)\ , s\_R(x\_2)\ , \dots\ , s\_R\ (x\_p))$ (16)

3- We find the correlation matrix U where

$u_{JK} = \frac{1}{4}(\mathrm{S}(Y_J + Y_K)^2 - \mathrm{S}(Y_J - Y_K)^2)$ (17)

$\mathrm{Y} = (Y_1, Y_2, \dots, Y_p)$

4- We find the matrix E of the characteristic vector to U and we find the following:

A) We find the characteristic vector V=YE

B) We find the robust variance $V=(V_1, V_2, \dots, V_p)$

Λ = diag $(s^2(V_1), \dots, s^2(V_p))$

C) We find $\mu =$ (Y) Em where m= $(m(V_1), \dots, m(V_p))'$,

$\hat{\Sigma}(\mathrm{Y}) = \mathrm{E}\Lambda E'$

So mean estimate $\hat{\mu}_{ORGK}$ (x) = D$\hat{\mu}$(Y) and variance

$\hat{\Sigma}_{ORGK}(\mathrm{x}) = \mathrm{D}\hat{\Sigma}(\mathrm{Y})D'$

3. Results

The research was applied to data that causes air pollution, which includes four types of gases (Methane CH_4 , Carbon monoxide CO, Nitrous gas NO_2 , Sulfur dioxide SO_2), from four stations for measured rates of monthly air pollution for 2019.

Genetic algorithms (MRCD, OGK) were applied, and the proposed algorithm was called (Orthogonalized Regularized Gnanadesikan – Kettenring) and is abbreviated as (ORGK), to find the most common types of gases that cause air pollution and compare the algorithms depending on the matrix of covariance and determinant.

Table 1 shows the results of finding the gases that most affect air pollution by comparing the covariance matrix for four stations by using the MRCD Genetic Algorithm as follows:

Table (1) : Show types of gases that cause air pollution by using MRCD

Station / gases	M1	M2	M3	M4
CH_4	0.5012	0.5088	0.5013	0.7572
CO	0.5222	0.5134	0.5022	0.6772
NO_2	0.5000	0.5000	0.5000	0.5000
SO_2	0.5001	0.5000	0.5000	0.5000
determinint	0.0654	0.0653	0.0630	0.1271

The results showed that the M3 station has the lowest determinant of the covariance matrix (0.0630), and by comparing variances, the least polluting gases are represented by (Nitrous NO_2) for all stations, as for the stations (M1,M2,M3), carbon monoxide (CO) is the most common cause of air pollution.

Table 2 shows the gases that cause pollution using (OGK) as follows:

Table (2) : Types of gases that cause air pollution by using OGK

Station / gases	M1	M2	M3	M4
CH_4	0.0741	10.7499	0.0701	6.5959e+03
CO	16.8829	17.4448	0.3508	9.9735e+03
NO_2	6.0474e-05	1.7000e-05	4.8988e-05	2.7352e-05
SO_2	1.8596e-04	1.3862e-04	1.1067e-04	1.4350e-05
determinint	5.5912e-10	3.0022e-07	3.1151e-11	0.0052

The result indicates that station $M3$ represents the lowest determinant of the covariance matrix (3.1151e-11) as it represents the station that measured the least air pollution, while the gas causing the pollution is carbon monoxide (CO) for all stations.

Table (3) : Types of gases that cause air pollution by using OGKR

Station / Gases	M1	M2	M3	M4
CH_4	0.0285	0.0209	0.0056	0.8716
CO	0.0402	0.0253	0.0041	0.3808
NO_2	0.0134	0.0196	0.0025	0.0902
SO_2	0.0081	0.0213	0.0022	0.1572
Determinint	2.1556e-08	1.8015e-07	1.5933e-11	0.0017

As for the results of applying the algorithm OGKR, the stations that recorded the least air pollution were the station $M3$, while the most pollution gases were represented by Methane gas CH_4 , the carbon monoxide gas by a slight difference. As for stations M1 and M2, it recorded the highest pollution of carbon monoxide gas. As for station 4 , Methane gas was the highest pollution, then gas carbon monoxide.

To compare the algorithms that were applied to find the most polluting gases, and through the results reached, this determinant of OGKR algorithm less than the determinant of OGK and MRCD algorithm

4. Conclusion

Based on the results reached through the application of genetic algorithms, carbon monoxide CO gas is considered the leading cause of air pollution resulting from car exhausts, as well as the operation of generators for many hours due to power outages, such as gas methane. CH_4 least polluting.

5 . References

[1] Clifford lam .High-Dimensional Covariance Matrix Estimation.Department of statistics.London School. http://stats.lse.ac.uk.

[2] Fatimah Abdul–Hammeed and Mohammad Huseen Abdul–Hammeed,(2011)," 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.2011),No1009.

[3] Fatimah Abdul – Hameed and Sabah Manfi (2016)," Compared with genetic algorithm Fast-MCD-Nested Extension and neural network multilayer Backpropagation", JOURNAL OF ECONOMIC&ADMINISTRATIVE SCIENCE ,(Jun.2016),No 22(89),pp.381-395.

[4] Gnanadesikan,R.and Kettection,j,(1972),"Roubust estimates,residuals,and outlier detection with multiresponse data.Biometrics", 28,pp. 81-124 .

[5] Hubert,M.and Debruyne,M,(2010),"Minimum Covariance Determinant.Wiley Interdisciplinary Reviews:Computional Statistics" ,2, 36-34.

[6] Jan Kalina,jurjen Duintjer Tebbens,and Anna Sehlenker .Robustness of High Dimensional Data Mining.https://www.sementisy scholar.org.

[7] Ledoit , O. and Wolf ,M,(2019),"Quadratic Shrinkage for Large Covariance Matrices ",University of Zurich .(November.2019).

[8] Ledoit , O. and Wolf ,M,(2004),"A well-conditioned estimator for large-dimensional covariance matrices", Journal of Multivariate Analysis,88(2),pp.365-411.

[9] Mia.Hert.and Paterj.Rousseeuw.and Tim Verdonck,(2011),"Adeterministic algorithm for robust Location and scatter",pp.journal of computational(2012).

[10] Peter j.Rousseeuw,Steven Vandumffel,Tim Verdonck,(2018),"The Minimum Regularized Covariance Determinant Estimator",ar Xiv:1701.07086v3(November .2018) ,29.

[11] Rousseeuw,P.and Van Driessen,K,(1999),"Afast algorithm for the Minimum Covariance Determinant estimator", Technometrics 41,pp. 212-223.

[12] Virgile Fritsch,Gael varoquaux,Thyreau Benjamin,Jean-Baptiste poline,Bertrand Thirion,(2011),"Detecting outlying subjects in High-Dimensional Neuroimaging Data sets with Regularized Minimum Covariance Determinant",HAL .(sep. 2011),27.

[13] Yilun,Ami wiesel,AlfredO,(2010),"HeroIII.Robust Shrinkage Estimtion of high Dimensional Covariance Matrices",arXiv:1009.5331v1[stat.ME].27(sep.2010).

[14] Zongliang Hu ,Kai Dong , Wenlin Dai and Tiejan Tong,(2017),"Acomparison of methods for Estimating the Determinent of High-Dimensional Covariance Matrix".(August.2017),16.

https://doi.org/10.31272/jae.i147.1309
https://admics.uomustansiriyah.edu.iq/index.php/admeco
P-ISSN: 1813–6729 E-ISSN: 2707-1359

تقدير مصفوفة التباين عالية الابعاد باستخدام الخوارزمية الجينية OGK

فاطمة عبد الحميد جواد البيرماني
قسم الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد ، بغداد ،العراق
Email: Fatimah.a@coadec.uobaghdad.edu.iq, ORCID ID:/ https://orcid.org/0000-0001-8767-1410

المستخلص

في هذا البحث ، تم تقدير مصفوفة التباين عالية الابعاد عندما تكون عدد المتغيرات اكبر من عدد المشاهدات ، فقد تم تطبيق الخوارزمية الجينية OGK لايجاد مصفوفة التباين حيث تم اقتراح تعديل للخوارزمية الجينية OGK بالاعتماد على معلمة التنظيم ρ ومصفوفة الهدف T واطلق عليها (Orthogonalized Regularized Gnanadesikan – Kettenring) يمكن كتابتها بشكل مختصر ORGK. اخذت البيانات من اربع محطات تمثل المعدلات الشهرية لمجموعة من الغازات الملوثة لمدة سنة .تم قياس المعدلات الشهرية لاربعة انواع من الغازات (غاز الميثان CH_4 ، غاز اول اكسيد الكاربون CO، غاز اكسيد النتروجين NO_2 ، غاز ثاني اكسيد الكبريت SO_2) . في هذه الدراسة تم اجراء مقارنة بين الخوارزمية الجينية OGK و ORGK و MRCD من خلال حساب محدد مصفوفة التباين وتحديد اكثر الغازات تلوثا . اظهرت النتائج ان غاز اول اكسيد الكربون COهو السبب الرئيسي للتلوث وان الخوارزمية الجينية ORGK المعتمدة على معلمة التنظيم ومصفوفة الهدف تأثيرا واضحا في الحصول على اقل محدد لمصفوفة التباين اطلق عليها (Orthogonalized Regularized Gnanadesikan – Kettenring) ORGK.

معلومات البحث

تواريخ البحث:

تاريخ تقديم البحث: 26 / 11 / 2024
تاريخ قبول البحث: 9 / 1 / 2025
عدد صفحات البحث 24 - 29

الكلمات المفتاحية:
معلمة التنظيم ، الحد الادنى لمحدد التباين المنتظم ، مصفوفة الهدف ، مهلنوبس

المراسلة:
أسم الباحث: فاطمة عبد الحميد جواد البيرماني
Email: Fatimah.a@coadec.uobaghdad.edu.iq