# Estimation of Covariance Matrices

**Nur Izyan Binti Mustafa Khalid[1] Zahayu Binti Md Yusof [2]**

[1*]Kulliyyah Muamalat and Management Sciences,
Universiti Islam Antarabangsa Sultan Abdul Halim Mu'adzam Shah,
09300 Kuala Ketil, Kedah, Malaysia.

[2]School of Quantitative Sciences,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia.

[*]Corresponding author: nurizyan@unishams.edu.my

**Abstract**

This article introduces covariance regression analysis for a *p*-dimensional response vector. The proposed method explores the regression relationship between the *p*-dimensional covariance matrix and auxiliary information. We study two types of estimators: maximum likelihood and ordinary least squares (OLS). Then, we demonstrate that these regression estimators are consistent and asymptotically normal. Furthermore, we obtain the high dimensional and large sample properties of the corresponding covariance matrix estimators. Simulation experiments are presented to demonstrate the performance of both regression and covariance matrix estimates. An example is analysed from the Gross Domestic Product (GDP) to illustrate the usefulness of the proposed covariance regression model.

**Key Words** Gross Domestic Product (GDP); Covariance Regression; maximum likelihood; Covariance Matrix Estimation; ordinary least squares (OLS)

## Introduction

One popular approach for bringing down the number of parameters is assuming that the covariance matrix is sparse. In the last few years, various sparsity constraints have been imposed on either Σ [1,1,2,3,4,5] its inverse Σ-[1,6,7,8] or its eigenvalues.[9] An alternative approach is considering a factor model.10,11

Although the above approaches are useful, they all require that $n \to \infty$ to assure the consistency of covariance estimators. To overcome this challenge, one can employ the commonly used structured covariance matrix models that involve one or a small number of parameters, such as compound symmetry, autoregressive (e.g., AR(1)), and moving average (e.g., MA(1)). However, neither the spare covariance approach nor the structured covariance approach can directly link the covariance estimator to the auxiliary information (e.g., explanatory variables, spatial information, and social network). This motivates us to explore a new avenue to estimate the covariance matrices.

There are many motivating examples, and we provide two here. In the area of empirical with responses to Gross Domestic Product (GDP), the covariance matrix of responses plays an important role for the economic development.[12,13,14] In addition, many researchers have shown that such a covariance matrix is affected by firms' fundamentals. [15,16,17] This suggests that the covariance matrix can be explained by its associated relevant explanatory variables. We next observe that, in the field of spatial data analysis, the responses are often collected from different geographical locations. It is not surprising that the responses located near each other are likely to be strongly correlated. Accordingly, spatial statistics attempts to explain the covariance structure of responses by their geographical locations.[18,19,20] Finally, in the context of social networks, responses can be deter-mined through human behaviors. Researchers also found that activities of the connected network users are likely to be correlated. This suggests that the movement of responses is affected by the users' social networks.[21,22,23] Hence, it is natural to estimate the covariance of responses via the social network structure.

Before proposing our covariance estimation method, we review two types of linkages between the covariance and auxiliary information (or covariates). The first type does not directly link the covariance to the auxiliary information. By using the fact that the mean vector of responses is a function of covariates, however, the resulting estimate of covariance is a function of covariates.[24,25] The second type directly links to the covariates under special model structure.[26,27,28] It is worth noting that Anderson24 also modeled as a linear combination of symmetric matrices, and later Szatrowski[25] and Zwiernik et al.[29] further studied the properties of the covariance estimates under the linear structure.

Inspired by the three motivating examples and the above methods for modeling the covariance, we integrate the similarity concept Johnson and Wichern,[30] the direct linkage approach, and Anderson's[24] linear combination method together, and then propose a co-variance regression model to directly quantify the relationship between the covariance and a linear combination of matrices induced by corresponding auxiliary information. We next present two types of estimators, the maximum likelihood estimator and the ordinary least squares demonstrate that those estimators are asymptotically normal. It is worth noting that the maximum likelihood estimator is computationally complex and the ordinary least squares estimator is inefficient.

## Maximum Likelihood

This method, however, show a sensitivity to various factors, such as violation of the normality assumption, presence of outliers and samples that show the effects of asymmetry and excess kurtosis compared to the multivariate normal distribution.

It is reasonable to assume that the presence of outliers effects can be explained by observing outliers in the sample and it is also relevant to consider sample size because large data sets are subject to a large number of these observations. Moreover, the violation of the assumption of normality can also be caused by these observations; therefore, a practical (but not always feasible) alternative is to apply a transformation to the data. [14] state that – depending on the transformation to be used – the relationships between variables may be nonlinear. However, applying this statement to structural equation models, where the nature of the relationships between the variables are linear, the application of a transformation may complicate the interpretation of results as well as affect the quality of model fit. [26]

In the statistical literature the word “robust” is synonymous with “good.” There are many classical statistical procedures such as least squares estimation for multiple linear regression and the t–interval for the population mean μ. A given classical procedure should perform reasonably well if certain assumptions hold, but may be unreliable if one or more of these assumptions are violated. A robust analog of a given classical procedure should also work well when these assumptions hold, but the robust procedure is generally tailored to also give useful results when a *single, specific assumption is relaxed.*

## Ordinary Least Squares (OLS)

These methods, however, show a sensitivity to various factors, such as violation of the normality assumption, and samples that show the effects of asymmetry and excess kurtosis compared to the multivariate normal distribution.[17]

This study is to apply robust statistical procedures to the regression credibility estimation, which are insensitive to the occurrence of outlier events in the data. A review of robust estimators that appeared in the literature is provided, including robust estimators that simultaneously attain maximum breakdown point and full asymptotic efficiency under normal errors.

## Methodology

In this study, we employ our proposed covariance regression model to analyze the quarterly returns of $p$ = 660 stocks in Malaysia Stock Market from 2010 to 2014, where the data were collected from the Trading Economics database. For each given quarter, the response variable $Y$ is the corresponding returns (in percentages) of the 660 stocks, standardized by subtracting the sample mean. There are $T$ = 20 quarters in total. In empirical finance, the covariance matrix of a large pool of stock returns measures the stock return comovement or synchronicity. As indicated by Roll,[15] stocks' comovement depends on the relative amounts of firm and market level information capitalized into stock prices, which is also directly related to the theory of market efficiency.[31] Since the pioneering work of Roll,[15] considerable effort has been devoted to exploring the relationship between the stock return comovement (or synchronicity) and firms' fundamentals, which motivates us to employ our proposed method to estimate the covariance of stock returns via some relevant information of firms' fundamentals.

In practice, common experience suggests that the returns of the stock in the same industry are more highly correlated than those of two stocks in different industries, which was confirmed by Chan et al..17 In addition, Chan et al.,[16,17] found that the cash flow, stock size, and book-to-market ratio can help to explain the covariation in returns. Furthermore, Gul et al.[32] employed leverage, size, and book-to-market ratio as control variables that are known to affect the stock return synchronicity. According to the above and extant literature, we consider the following *k*= 5 covariates to represent firms' fundamentals in this study: IND (industry); LEV (leverage computed by liability-to-asset ratio); CF (cash flow of the firm); SIZE (measured by the logarithm of market value); and BM (book-to-market ratio). We label them as covariates *X*(*k*) = (*X*1*k;; Xpk*)T *2* R*p* for *k* = 1*; ;* 5, respectively. For the variable IND, let the off-diagonal element in the associated similarity matrix be 1 if two stocks belong to the same industry, and 0 otherwise, keeping this setting across all 20 quarters. For each given quarter, we next standardize the rest of the four variables via *p* = 660 observations so that they have zero mean and unit variance. Subsequently, we set the off-diagonal elements of the similarity matrices to be exp *f* (*Xj*1*k Xj*2*k*)2*g* for stocks *j*1 ≠*j*2 and covariates *k* = 2*;* 5, and let the diagonal elements be zeros.

The goal of this study is to assess the performance of portfolio by solving the Markowitz optimization problem (12). To this end, we adopt the commonly used rolling window procedure (33:34:35) with the window length *n* = 1 to construct and assess portfolio returns. Suppose that the *t*-th quarter data is (*Yt; Xt*), where *Yt 2* R*p* 1, *Xt* = (*Xt*(1)*; ; Xt*(*K*)) *2* R*p K* and *t* = 1*; T* . Since the covariance matrix is time varying, we utilize each single period data at time *t* to t the proposed covariance regression model and then estimate the covariance matrix *t* = Cov(*Yt*). Hence, the estimation is based on the sample size *n* = 1 and *p* = 660.

## Result and Discussion

TABLE I. Comparison of MLE and OLS covariance matrix estimates. Four measures are considered: the averaged execution time (Time, in seconds), the averaged spectral norm and Frobenius norm estimation errors (Spectral-Error and Frobenius-Error), and the percentage of the unconstrained covariance estimate being identical to its associated constrained estimate (Percentage). The response variable follows the normal distribution and the similarity matrices are *Wk*.

a) average execution time (Time, in seconds),
b) average spectrum norms and Frobenius norm estimation errors (Spectral Errors and Frobenius Errors),
c) and the percentage of uncontrolled covariance estimates is equal to the relevant constraint estimates (Percentage).

It was found that the equation matrix formed was Wk and the variables were found to be normally distributed.

**TABLE I: COMPARISON OF MLE AND OLS COVARIANCE MATRIX ESTIMATES.**

| | | *p = 50* | *p = 100* | *p = 200* | *p = 500* |
|---|---|---|---|---|---|
| MLE | Time | 10.4524 | 121.0258 | 1,648.1558 | 87,417.6100 |
| | Spectral-Error | 4.2083 | 3.0879 | 2.1557 | 1.3545 |
| | Frobenius-Error | 1.5945 | 1.0921 | 0.7761 | 0.4849 |
| OLS | Time | 0.0001 | 0.0004 | 0.0014 | 0.0205 |
| | Spectral-Error | 4.4538 | 3.2327 | 2.3444 | 1.4735 |
| | Frobenius-Error | 1.6677 | 1.1293 | 0.8326 | 0.5179 |
| | Percentage | 91.8% | 96.4% | 99.6% | 99.9% |

## Concluding Remarks

In short, by using the two proposed estimation methods, namely MLE and OLS along with theoretical properties, this can indirectly identify performance and support theoretical findings based on the covariance matrix. In addition, this article also suggests that these two types of estimators be used to analyze portfolio returns. The key component method is a tool for determining the main axis of propagation in a data set and making it easy to explore key data variables. The method used correctly is one of the most powerful in a set of data analysis tools.

In this paper, we utilize auxiliary information and employ a covariance regression approach to estimate the covariance matrix. Three estimation methods (MLE and OLS) have been proposed and their theoretical properties for both regression and covariance estimators are obtained. Simulation results demonstrate their performance, which supports theoretical findings. We also provide recommendations for using these three type estimators. An application for analyzing portfolio returns shows our proposed method performs well.

### REFERENCES

1. J. Z.Huang, N. Liu, M. Pourahmadi, and L. Liu,Biometrika. **93**. 85-98(2006).
2. P. J.Bickel, andE. Levina, The Annals of Statistics. **36.** 2577-2604. (2008a).

3. P. J.Bickel, andE. Levina, The Annals of Statistics. **36.** 199-227. (2008b)
4. T.T. Cai, and Liu, Journal of the American Statistical Association, **106**, 672-684. (2011).
5. C.Leng, and W.Li, Biometrika. **98**. 821830.(2011).
6. A. P. Dempster,Biometrics. **28**. 157-175.(1972).
7. N. P.Freidman, Current Direction in Physiological Sciences. **21.** (2008)
8. T. T.Cai, W. Liu, andX. Luo, Journal of the American Statistical Association. **106.** 594-607. (2011)
9. I. M. Johnstone, and A. Y.Lu, Journal of the American Statistical Association. **104**. 682-693. (2009)
10. J.Fan, Y. Fan, and J.Lv,Journal of Econometrics.**147.** 186-197. (2008).
11. J.Fan, Journal of Economics. **31.** (2011)
12. H. Markowitz,The Journal of Finance.**7.**77-91.(1952).
13. R. Jagannathan, and T.  Ma, The Journal of Finance. **58.** 16511684. (2003).
14. R. Kan, and G. Zhou, Journal of Financial and Quantitative Analysis. **42.** 621-656.(2007).
15. R. Roll, Journal of Finance. **43.** 541-566.(1988).
16. L. K. Chan, J.Karceski, and J. Lakonishok, Journal of Financial and Quantitative Analysis. **33.** 159-188.(1998).
17. L. K.Chan, J. Karceski, and J. Lakonishok, Review of Financial Studies.**12.** 937-974. (1999).
18. N. Cressie, *Statistics for Spatial Data* (Wiley Series in Probability and Statistics.1991).P. 301.
19. R. S. Bivand, E. J. Pebesma, and V. Gomez-Rubio, Springer. **25.** 135-141. (2008).
20. N. Cressie, and C.K. Wikle, *Statistics for spatio-temporal data* (John Wiley & Sons, Inc. 2011). P. 118
21. E. L. Glaeser, B. Sacerdote, and J. A. Scheinkman, The Quarterly Journal of Economics.**111.**507-548.(1996).
22. G. A.Akerlof, Econometrica.**65**. 1005-1027.(1997).
23. W. A. Brock, and S. N. Durlauf, The Review of Economic Studies.**68**. 235-260. (2001).
24. T. W. Anderson, The Annals of Statistics. **1.** 135 - 141(1973).
25. T. H. Szatrowski, The Annals of Statistics.**8.** 802-810.(1980).

26. R. L. Prentice, Biometrics.**44.** 1033-1048.(1988).
27. E. Demidenko. *Mixed models: Theory and applications*. (John Wiley and Sons, Inc. 2004), P. 221.
28. P. D. Hoff, andX. Niu,StatisticaSinica.**22.**729-753.(2012).
29. P.Zwiernik, C.Uhler, and D. Richards, Journal of the Royal Statistical Society. **79.** 1269-1292 (2014)
30. R. A. Johnson, andD. W. Wichern. *Applied multivariate statistical analysis* (Prentice Hall, 1992), P. 319.
31. Fama,The Journal of Finance. **25.** 383-417. (1970).
32. F. A. Gul, J. B. Kim, andA. A. Qiu, Journal of Financial Economics.**95.** 425-442.(2010).
33. Zivot and Wang, Biometrics. **65.** 1068-1077. (2009).
34. D. M.Xue, Biostatistics. **10.** 515-534. (2012).
35. J.Fan, Multivariate Analysis. **99.** 1015-1034. (2014).