ISSN: 2321-8819 (Online) 2348-7186 (Print) Impact Factor: 1.498 Vol. 5, Issue 12, December 2017

Appraisals on Resolving Singularity Problem of Covariance matrix in High Dimension

Hafeez Ahmed, Maznah Mat Kasim, Malina Zulkifli
School of Quantitative Science,
Universiti Utara Malaysia

Abstract: This is an expository essay that reviews the recent developments on resolving the singularity problem in the variance/covariance matrix in high dimension. Furthermore, the interrelated and multidimensional linear relationships between columns or rows in the covariance matrix structure result into zero determinants in its computations and are singular. This also means that the inverse of the underlying matrix becomes inflexible, that leads to a computational problem in the likelihood. Therefore, this paper intent to review some of the literatures in resolving the singularity problems and also to evaluate some of the methods used. This study is very vital since it will resolve major problems in many areas such as functional magnetic resonance imaging analysis of gene expression arrays, risk management and portfolio allocation.

Keywords: Multivariate analysis, High dimension, Covariance structure and Singularity problem.

1.0 Introduction

In multivariate analysis estimation, the concentration of distribution in the lower dimension subspace is a problem in the variance matrix structure thus, result into the singularity problem. Consequently, the estimation of the matrix becomes a difficult as the determinant and inverse of the sample covariance matrix are regularly needed in the likelihood estimation. However, high dimensional variance structure is very important in multivariate analysis method. In other words the collection of large vital statistical procedures, such as the principal component analysis(Jolliffe & Cadima, 2016); quadratic discriminant and linear analysis(Stewart, Ivy & Anslyn, 2014); regression and clustering analysis (Anderberg, 2014)involve the understanding of the variance structure.

The estimation of sample covariance matrix from high dimensional covariance matrix structure that involve the inverse and exactnessof the matrix, is suitable for resolving major problem in many areas such as functional magnetic resonance imaging(Fan, Han & Liu, 2014); analysis of gene expression arrays (Engel, Buydens & Blanchet, 2017); risk management and portfolio allocation(Dai, Lu & Xiu, 2017).

Additionally, the sample covariance matrix as the standard and most natural estimator behaves very badly and results into unacceptable decisions in high dimension situations. Assuming that, $\frac{p}{n} \rightarrow c \, \epsilon(0, \infty)$ is the sample covariance matrix largest eigen value that is not a reliable estimate of the population covariance matrix largest eigen value, whereby sample covariance matrix eigenvectors are closely orthogonal (Johnstone, 2001;Bickel, & Levina, 2008;Fan, Liao & Mincheva, 2013;Paul & Wang, 2016;Lee & Schnelli, 2016andLi, Wang & Yao, 2017).

On the whole in high dimension, the sample covariance matrix cannot be inverted, because of the scarcity nature of the structure, it is however difficult to apply it in many areas that needed the estimation of the exactnessof the matrix. In other to resolve this problem of singularity in high dimension the assumptions of covariance structure are needed in order to estimate the covariance or exactness of the matrix regularly. In recent time, many methods to resolve the singularity problem challenges have been proposed the most important ones are the eigen value, singular value, and Cholesky decompositions (Kazem & Hatam, 2017; Lan, Zhang, Ge, Cheng, Liu, Rauber & Zha, 2017). In high dimension the structural nature of covariance are frequently observed in many mathematical sciences from genomic data to statistical calculations to linear algebra and to financial analysis(Kuismin, Kemppainen & Sillanpaa, 2017).

In multivariate statistical analysis, the main worry is either the hypothesis testing or the inference of population parameters that are based on mostly high dimensional matrices where the covariance are classically used in the related terms(Kutner et al., 2005 and Dai, Wang, Xiong & Jiang, 2018). Alternatively, in genomic data within the microarray studies (Puccio, Grillo, Licciulli, Severgnini, Liuni, Bicciato & Peano, 2017). We need to deal with the concentrations from hundreds of genes at the same time in which the mean and variance estimates of genes are defined by high dimensional matrices. When working on such basic high dimensional covariance structures, we are face with the singularity problem (Liu, Maljovec, Wang, Bremer & Pascucci, 2017).

2.0 The Sample Covariance Structure

The interrelated and multidimensional linear relationships between columns or rows in the

covariance matrix structure result into zero determinants in its calculation are singular. This also means that the inverse of the underlying matrix becomes inflexible, that leads to a computational problem in the likelihood (Davoudi, Ghidary & Sadatnejad, 2017).

The estimation oflarge covariance and precision inverse matrix isa critical problem in modern multivariate analysis in many fields, ranging from economics and finance to biology, social networks, and health sciences (Fan et al., 2014a). When the dimension of the covariance matrix is large, the estimation difficulties usually challenging. Additionally, collection of substantial estimation errors can make significant different influences on the estimation correctness. Therefore, estimating large covariance and precision matrices attracts rapidly growing research attentions in the past period. In recent years researchers have proposed various regularization methods to reliably estimate large covariance and precision matrices in order to resolve the singularity problem in high dimension.

In other to estimate large covariance matrix the key assumption made in the literature is the scarcity of the target matrix of interest with many zero entries or nearly so (Bickel and Levina, 2008; Lam and Fan, 2009; El Karoui, 2010; Rigollet and Tsybakov, 2012). Moreover, in estimating the large precision matrices, it is frequent that the precision matrix is sparse, and the commonly used method for estimating the sparse precision matrix is to employ an ℓ-penalized maximum likelihood (Banerjee et al., 2008; Yuan & Lin 2007; Friedman et al., 2008; Rothman et al., 2008).

Additionally, to further moderate bias estimation, Lam and Fan (2009); Shen et al. (2012) suggested the estimation of non-convex penalties for sparse precision matrix and considered their hypothetical properties. In other to know more on the general theory of penalized likelihood methods see(Fan & Li, 2001; Fan & Peng, 2004; Zou, 2006; Zhao & Yu, 2006; Bickel et al., 2009 andWainwright,2009). Furthermore, for better explanation of this concept, the estimation that are based on robust estimates has been extended more on regularized rank-based methods (Liu et al., 2012a; Xue and Zou, 2012).

The rank-based method is mainly interesting when data are generated based on the process of non-Gaussian and heavy-tailed distributions in financial data (Han & Liu, 2013; Wegkamp & Zhao, 2013; Mitra & Zhang, 2014). This heavy-tailed data has been widely used in financial data analysis that are often modelled with the family of elliptical distributions (Hamada & Valdez, 2004;Sun, Frees& Rosenberg, 2008and Frahm & Jaekel, 2008).

3.0 Literature Review

3.1 Estimation of Variance-Covariance Matrix in High dimension

The estimation of high dimensional covariance matrix has becomes major problems in multivariate analysis, this find its applications in many areas, ranging from health sectors, finance and stock to social networks, and biology (Fan et al., 2014). The estimation of covariance matrix has general challenges, when the dimension, p is large as compare to the sample size, n, where the estimation of the likelihood ratio test is difficult to estimate, this lead to singularity problem of estimating the determinant and inversion. Moreover, it is well known that the observed data in sample covariance matrix is singular when $p > n$ situation. Additionally, the collections of huge quantity of estimation errors will make significantly contrary effects on the estimation precision.

In the past decade estimating high dimension covariance matrix has involved fast increasing research considerations. However, in current years researchers have suggested various regularization methods to reliably estimate large covariance and precision matrices. The main assumptions in the estimation of high dimension covariance matrix made in previous study is that the population covariance matrix of interest is sparse, with many entries are zero or close to it (Bickel and Levina, 2008; Lam and Fan, 2009;El Karoui, 2010; Rigollet and Tsybakov, 2012).

On the other hand, to estimate the high dimension sample covariance matrices, it is frequently the situation that the sample covariance matrix is sparse. A frequently used method for estimating the sparse sample covariance matrix is to employ an ℓ-penalized maximum likelihood in Banerjee et al. (2008); Yuan and Lin (2007); Friedman et al. (2008); Rothman et al. (2008).In addition to further reduce the estimation bias in high dimension, Lam and Fan (2009); Shen et al. (2012) suggested non-convex penalties for sparse sample covariance matrix estimation and studied their theoretical properties. For more general theory on penalized likelihood methods, see Fan and Li (2001); Fan and Peng (2004); Zou (2006); Zhao and Yu (2006); Bickel et al. (2009); Wainwright(2009).

3.2 The High dimensional Problems in Covariance Matrix Structure

In multivariate analysis, the high dimension covariance matrix structure poses many difficulties

to the applications of statistical theory, techniques and implementations in those problems. For instance, in linear regression model with noise variance σ^2, when the dimensionality p is large compare to the sample size n, ordinary least squares (OLS) estimator is not well performed and conditioned due to the singular matrix nature. This indicates two familiar singularity problems in high dimensional estimation, co linearity or false correlations and the noise build-up. The false correlations among the predictors are an inherent problem in high dimensional estimation of the covariance matrix.

Additionally, two main causes of co linearity can be seen in the population and the sample levels. There can be high false correlation even for independent and identically distributed (*i.i.d.*) covariance matrix when p is larger than n (Fan &Lv, 2008; Fan &Lv, 2010 and Fan et al, 2010). A different challenging case is the data loading problems in high dimensional structure in Hall et al., 2005.

The presence of high co linearity in covariance matrix structure the issues of over fitting and estimation identifications occur where noise build-up is a common occurrence in high dimensional estimation. The estimation of covariance matrices of high-dimensional is a crucial problem, and it arise in various applications such as machine learning (Fan &Lv, 2008), microarray (Yu, Feng, Miller, Xuan, Hoffman, Clarke & Wang, 2010), finance and stocks market (Park & O'Leary, 2010), functional magnetic resonance imaging, risk management and portfolio allocation (Ding, Sun, Sun, Chen, Zhou, Zhuang& Du, 2014).Most of the researchers has been paying attention to the estimation of $p \times p$ covariance matrix, $\boldsymbol{\Sigma}$ and its inversion, $\boldsymbol{\Sigma}^{-1}$ also known as the precision matrix. As the number of variable, p approached the sample size, n the determinant of sample covariance matrix, $|\boldsymbol{S}|$ become nearly singular.

3.3 How Resolving the Singularity Problem in High dimension

In high dimension on how to resolve the singularity problem is one of the common tasks in the statistical computations. Furthermore, there are number of methods that exist to solve this problem. It is therefore important to review some of the recent methods used in resolving the singularity problems.

3.3.1 Eigenvalue Decomposition in Resolving Singularity Problem

The basic details about eigen value decomposition to resolve the singularity problem in high dimension covariance matrix is being accessed from a geometrical angle, where the eigenvectors shows the direction of pure stretch and the eigen values the extent of stretching. Most matrices are complete with complex eigenvectors and form basis of the basic vector space. On the whole important class are the symmetric matrices, whose eigenvectors form an orthogonal basis of $\mathbb{R}^n$. A non-square matrix A does not have eigen values. In their place, one uses the square roots of the eigen values of the associated square positive semi-definite matrix$K = A^T A$, that are refers to as the singular values of the unique matrix. However, on how to compute the eigen values and eigenvectors is cumbersome.

Given,$A$is a symmetric matrix, Urepresents a matrix with the set of eigenvectors of A, where the main diagonal matrix $\wedge$are the eigen values of A, then the diagonal elements are written with the following equality for every A, U, and $\wedge$.

$$AU = U \wedge \quad (2)$$

However, the decomposition of the eigen value of Acan be set up by,

$$A = U \wedge U^{-1} \quad (3)$$

Consequently, we compute the square root of A by taking the square root of $\wedge$through

$$A^{-1} = U \wedge^{1/2} U^{-1} \quad (4)$$

If A is positive semi-definite matrix, the eigen value decomposition of this matrix always exists and the related eigen values are always positive or zero. The example of this type of matrix can be expressed by the correlation, covariance, and cross-product matrices (Healy, 1986). The decomposition of the structural matrix can disappear if the unique matrix is singular and there is always the need to convert this matrix into non-singular form by the dimensional reduction method (Johnson & Wichern, 2002; Rencher, 2002).

3.3.2 Singular Value Decomposition in High dimension

The singular decomposition methods in high dimension is represented by a data set with n measurements on $p$dimensions is given by an$n \times p$ data matrix X. In highdimensionalsettings where p is large, it is often required to work with a low-rank estimate of the data matrix. The most widespread low-rank estimate is the singular value decomposition (SVD).

Given X, an $n \times p$data matrix, the *SVD* factorizes Xas $X = UDV'$, where $U \epsilon \mathbb{R}^{n \times n}$and $V \epsilon \mathbb{R}^{p \times p}$are orthogonal matrices and $D \epsilon \mathbb{R}^{n \times p}$are always zero

except on its main diagonal with diagonal entries in shrinking order. The best rank K eigen value estimate to $X, \hat{X}_K$, in the Frobenius and operator norms are always given by the first K eigen value, right singular vectors and singular values of the *SVD*:

$\hat{X}_K = \sum_{k=1}^{K} d_k u_k v'_k$. The *SVD* of X is also closely related to the eigen decomposition of $X'X$. In detail, if UDV' is an *SVD* of X, then $V(D'D)V'$ is an eigen decomposition of $X'X$. Thus, the eigen values of $X'X$ are the squares of the singular values of X, and the eigenvectors of $X'X$ are the right singular vectors of X. Another method of resolving the singularity problem is the method of Cholesky decomposition that is similar to the singular value decomposition.

3.3.3 Cholesky Decomposition in High dimension

In high dimension covariance matrix the Cholesky factor is the decomposition of a positive definite matrix into the product of a lower triangular matrix and its conjugate transpose for an efficient algebraic solutions. In specific term one can obtain a positive definite banded estimator of the covariance matrix at the same computational cost as the popular banded estimator proposed by Bickel and Levina (2008b), which is not guaranteed to be positive definite.

Let A denotes a symmetric and positive definite matrix, the Cholesky decomposition of A can be found by an upper triangular matrix U having severely positive diagonal entries such that

$$A = U^T U$$

Now, the matrix U referred to square root of A. We take A also to be symmetric, then $U^T = U$, where, $A = UU$. But if A is positive semi-definite, i.e. some eigen values are zero, we use a numerical tolerance in the decomposition of A In this method, similar to its previous alternatives, it cannot preserve the unique structure of the matrix when the singularity problem is solved by this decomposition and a new non-singular matrix is defined under the unique dimension of A (Johnson & Wichern, 2002; Rencher, 2002).

3.0 Conclusion and Summary

The problem of singularity was identified as a major concern in high dimension covariance matrix structure as challenging. The review on some current development in high dimension covariance matrix structure has reveal some of the methods that was used and suggested.

Also, the importance of covariance matrix estimation was highlighted in some practical situations.

4.0 References

Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks* (Vol. 19). Academic press.

Bickel, P. J., & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 2577-2604.

Dai, C., Lu, K., & Xiu, D. (2017). Knowing Factors or Factor Loadings, or Neither? Evaluating Estimators of Large Covariance Matrices with Noisy and Asynchronous Data.

Dai, W., Wang, S., Xiong, H., & Jiang, X. (2018). Privacy Preserving Federated Big Data Analysis. In *Guide to Big Data Applications* (pp. 49-82). Springer International Publishing.

Davoudi, A., Ghidary, S. S., & Sadatnejad, K. (2017). Dimensionality reduction based on distance preservation to local mean for symmetric positive definite matrices and its application in brain–computer interfaces. *Journal of Neural Engineering*, *14*(3), 036019.

Engel, J., Buydens, L., & Blanchet, L. (2017). An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics. *Journal of Chemometrics*, *31*(4).

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(4), 603-680.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 295-327.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, *374*(2065), 20150202.

Kazem, S., & Hatam, A. (2017). A modification on strictly positive definite RBF-DQ method based on matrix decomposition. *Engineering Analysis with Boundary Elements*, *76*, 90-98.

Kuismin, M. O., Kemppainen, J. T., & Sillanpää, M. J. (2017). Precision Matrix Estimation with ROPE. *Journal of Computational and Graphical Statistics*, (just-accepted).

Lan, L., Zhang, K., Ge, H., Cheng, W., Liu, J., Rauber, A., ...& Zha, H. (2017). Low-rank decomposition meets kernel learning: A generalized Nyström method. *Artificial Intelligence*.

Lee, J. O., & Schnelli, K. (2016). Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *The Annals of Applied Probability*, *26*(6), 3786-3839.

Li, Z., Wang, Q., & Yao, J. (2017). Identifying the number of factors from singular values of a large sample auto-covariance matrix. *The Annals of Statistics*, *45*(1), 257-288.

Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2017). Visualizing high- dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, *23*(3), 1249-1268.

Paul, D., & Wang, L. (2016). Discussion of "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation". *Electronic Journal of Statistics*, *10*(1), 74-80.

Puccio, S., Grillo, G., Licciulli, F., Severgnini, M., Liuni, S., Bicciato, S., ...& Peano, C. (2017). WoPPER: Web server for Position Related data analysis of gene Expression in Prokaryotes. *Nucleic Acids Research*.

Stewart, S., Ivy, M. A., & Anslyn, E. V. (2014). The use of principal component analysis and discriminant analysis in differential sensing routines. *Chemical Society Reviews*, *43*(1), 70-84.

Sun, J., Frees, E. W., & Rosenberg, M. A. (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics*, *42*(2), 817-830.