

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Contributions to high-dimensional data analysis: some applications of the regularized covariance matrices

A thesis submitted in partial fulfilment of the requirements for the

degree of

Doctor of Philosophy

in

Statistics

AT MASSEY UNIVERSITY, ALBANY
NEW ZEALAND.

Insha ULLAH

March 2015

Abstract

High-dimensional data sets, particularly those where the number of variables exceeds the number of observations, are now common in many subject areas including genetics, ecology, and statistical pattern recognition to name but a few. The sample covariance matrix becomes rank deficient and is not invertible when the number of variables are more than the number of observations. This poses a serious problem for many classical multivariate techniques that rely on an inverse of a covariance matrix. Recently, regularized alternatives to the sample covariance have been proposed, which are not only guaranteed to be positive definite but also provide reliable estimates. In this Thesis, we bring together some of the important recent regularized estimators of the covariance matrix and explore their performance in high-dimensional scenarios via numerical simulations. We make use of these regularized estimators and attempt to improve the performance of the three classical multivariate techniques in high-dimensional settings.

In a multivariate random effects models, estimating the between-group covariance is a well known problem. Its classical estimator involves the difference of two mean square matrices and often results in negative elements on the main diagonal. We use a lasso-regularized estimate of the between-group mean square and propose a new approach to estimate the between-group covariance based on the EM-algorithm. Using simulation, the procedure is shown to be quite effective and the estimate obtained is always positive definite.

Multivariate analysis of variance (MANOVA) face serious challenges due to the undesirable properties of the sample covariance in high-dimensional problems. First, it suffer from low power and does not maintain accurate type-I error when the dimension is large as compared to the sample size. Second, MANOVA relies on the inverse of a covariance matrix and fails to work when the number of variables exceeds the number of observation. We use an approach based on the lasso regularization and present a comparative study of the existing approaches including our proposal. The lasso approach is shown to be an improvement in some cases, in terms of power of the test, over the existing high-dimensional methods.

Another problem that is addressed in the Thesis is how to detect unusual future observations when the dimension is large. The Hotelling T^2 control chart has traditionally been used for this purpose. The charting statistic in the control chart rely on the inverse of a covariance matrix and is not reliable in high-dimensional problems. To get a reliable estimate of the covariance matrix we use a distribution free shrinkage estimator. We make use of the available baseline set of data and propose a procedure to estimate the control limits for monitoring the individual future observations. The procedure do not assume multivariate normality and seems robust to the violation of multivariate normality. The simulation study shows that the new method performs better than the traditional Hotelling T^2 control charts.

Acknowledgements

My Ph.D project was supported by Higher Education Commission (HEC) of Pakistan. I gratefully acknowledge the financial support of the HEC.

I am grateful to my supervisor, Dr. Beatrix Jones for being my supervisor. She gave me the opportunity to learn from her and work together with her. Indeed, it is because of her, I managed to accomplish the task. Thanks Beatrix for your constant support, patience, suggestions and constructive criticisms. You made it a fruitful journey for me. I would like to thank my co-supervisor Prof. James Curran for his valuable feedbacks.

I have greatly benefited from the group of statistician working at the Institute of Natural and Mathematical Sciences, Massey University. I would like to thank all members of the group.

I also acknowledge the support of my family and friends. Thanks to Hoor and Hamood for being patient with me.

Contents

Abstract	i
Acknowledgements	iii
List of Figures	vi
List of Tables	xiii
1 Introduction	1
1.1 Introduction	1
2 Regularized estimation of high-dimensional covariance matrices	5
2.1 Introduction	5
2.2 Moore-Penrose generalized inverse	6
2.3 Shrinkage Estimate	7
2.3.1 Computation of the shrinkage intensity ρ	8
2.4 Penalized Normal Likelihood	10
2.4.1 Ridge Regularization	11
2.4.1.1 Selection of the ridge parameter ρ	11
2.4.2 Lasso Regularization	12
2.4.3 Adaptive lasso and SCAD penalty functions	14
2.4.4 Choosing the optimal value of the penalty parameter ρ	16
2.5 Numerical simulations	18
2.6 Summary and conclusion	22
2.7 Contributions of the Chapter	23
3 Hierarchical covariance estimation	28
3.1 Introduction	28
3.2 Hierarchical covariance estimation via EM algorithm	31
3.3 Applications	33
3.3.1 Numerical experiments	34
3.3.2 Real data example	45
3.4 Conclusion	48
3.5 Contributions of the Chapter	49

4	Regularized MANOVA for high-dimensional data	50
4.1	Introduction	50
4.2	Previous works	53
4.3	LASSO Regularization	56
4.3.1	Selection of ρ	56
4.4	Simulation study	57
4.4.1	Simulation results	59
4.5	Practical application	66
4.6	Conclusion and discussion	69
4.7	Contributions of the Chapter	70
5	Monitoring future observations in the high-dimensional setting	72
5.1	Introduction	72
5.2	Shrinkage estimate of a covariance matrix	75
5.3	Proposed procedure	80
5.4	Simulation study	82
5.5	The in-control run length performance	86
5.6	Practical applications	87
5.6.1	Example 1: Gene expression data	87
5.6.2	Example 2: Chemical process data	92
5.7	Summary and Conclusion	96
5.8	Contributions of the Chapter	97
6	General Conclusions	100
6.1	Regularized estimation of the high-dimensional covariance matrices	101
6.2	Hierarchical covariance estimation	101
6.3	Regularized MANOVA for high-dimensional data	102
6.4	Monitoring individual future observations in the high-dimensional setup	103
6.5	Commonalities and future work	104
A	Supplementary Figure for Chapter 4	106
B	Supplementary Figure for Chapter 5	109
C	DRC forms	118
	References	119

List of Figures

2.1	A schematic diagram showing the regularized estimates against the maximum likelihood estimates for the elements of the inverse covariance matrix (a) the ridge, (b) the lasso, (c) the adaptive lasso, (d) the SCAD regularized estimates.	16
2.2	Off-diagonal elements of simulated correlation matrices for $p = 20$ using the algorithm of Schäfer & Strimmer (2005a). The proportion of non-zero off-diagonal elements in step 2 of the algorithm is (a) 20%, (b) 30%, and (c) 40%. For a fixed value of p as we increase the proportion of the non-zero elements in the off-diagonal positions, the size of the off-diagonal elements in the correlation matrix decreases.	21
2.3	Ordered eigenvalues of a true and estimated covariance matrices. A true covariance matrix is generated using the random method with proportion of non-zero off-diagonal elements equals to 50% and $p = 40$. The covariance matrix is estimated 1000 times, using 1000 samples for each $n \in \{20, 40, 1000\}$ from a multivariate normal distribution and the average eigenvalues of the estimated covariance matrices are presented. The diagonal elements of the estimated covariance matrices are not penalized in any of the regularization procedures.	24
2.4	An exchangeable covariance structure is used with $b = 0.6$. The box-plots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods. As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.	25

2.5	A random covariance structure is used with the proportion of non-zero edges equals to 50%, 30%, 20%, and 10%, respectively, for p equals to 10, 20, 40, and 80. The box-plots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods. As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.	26
2.6	An AR(1) covariance structure is used with $b = 0.6$. The box-plots show the distributions of the three loss functions for the five competing procedures. The three loss functions are calculated for each of 1000 samples of size $n = 20$ from a multivariate normal distribution with $p \in \{10, 20, 40, 80\}$. Since the shrinkage, lasso, adaptive lasso, and SCAD regularization allow to penalize the diagonal, the diagonal elements in (a) are left unpenalized while in (b) they are penalized for all the four methods. As ridge regularization does not allow to penalize the diagonal elements; therefore, the distributions in (b) are the replicates of the results in (a) for ridge regularization.	27
3.1	Simulated data on first two principle components for the three cases (a) easiest, (c) moderate, and (e) hard.	35
3.2	Typical cross-validation scores obtained over the first five iterations of the proposed EM algorithm. We use 100% non-zero off-diagonal elements in \mathbf{U}^{-1} and \mathbf{C}^{-1} . The elements of \mathbf{U} are of a moderate size. In (a) $p = 5, m = 10, r = 5$ and in (b) $p = 15, m = 10, r = 5$	37
3.3	Estimates of different elements of the between-group covariance matrix in an easy case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of five diagonal elements and the horizontal line in each panel represents the true value.	39
3.4	Estimates of different elements of the between-group covariance matrix in a moderate case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of the five diagonal elements and the horizontal line in each panel represents the true value.	40
3.5	Estimates of different elements of the between-group covariance matrix in a hard case. Each box-plot is made up of 1000 estimates of the same element using 1000 different data sets. The gray boxes represent the estimate of five diagonal elements and the horizontal line in each panel represents the true value.	41

3.6 The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use \mathbf{C} as in (3.17) and \mathbf{U} in (3.15) is scaled to the easy case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.1. 43

3.7 The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use \mathbf{C} as in (3.17) and \mathbf{U} in (3.15) is scaled to the moderate case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.2. 44

3.8 The distributions of the estimated eigenvalues of a between-group covariance matrix with 1000 different data sets. We use \mathbf{C} as in (3.17) and \mathbf{U} in (3.15) is scaled to the hard case: (a) $m = 10$ and $r = 5$, (b) $m = 10$ and $r = 20$, (c) $m = 30$ and $r = 5$, and (d) $m = 30$ and $r = 20$. See the mean squared errors of the estimated eigenvalues in Table 3.3. 45

3.9 First two principle components of glass chemical composition data. 48

4.1 Distribution of the sum of absolute errors $(\sum_{i=1}^p |\hat{\lambda}_i - \lambda_i| / \sum_{i=1}^p \lambda_i)$ in the estimated eigenvalues under (a) exchangeable and (b) AR(1) covariance structures both with $b = 0.6$. For each value of $p \in \{10, 20, 40, 80\}$, 1000 samples of size 20 are simulated from a multivariate normal distribution. Eigenvalues of the true covariance matrix are estimated using shrinkage, ridge, and lasso regularization and the sum of absolute errors are calculated for each of the 1000 samples. Note that, the estimation error increases as we increase p and the lasso regularization maintains the highest accuracy. 61

4.2 Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05. 62

4.3 Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.8$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05. 63

4.4 Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05. 64

4.5 Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.7$. For each value of $p \in [2, 30]$, the point on a power curve is the average of 1000 experiments. The significance level is 0.05. 65

4.6	Time comparison of 3 competing regularization procedures under two different covariance structures: exchangeable and AR(1). Each point in the graph is averaged over 10 replicates. The covariance structure does not make big difference in computational time for ridge and shrinkage therefore only shown for exchangeable. The computational time for principal components and and generalized inverse is not shown but lie below the shrinkage estimate.	66
4.7	Serial correlation coefficients of soil compaction at shallower depths with soil compaction at all the deeper depths after adjusting for group means. Each sequence of joined points of the same color represents the correlation of the measurements at a certain depth with the measurements at deeper levels (with the depth value given on the x axis). Note that for each of the 18 depths (variables) we have 21 observations.	68
4.8	Projection of the data (variables are standardized to have zero means and unite standard deviations) onto the first two principal components of the data after adjusting for group means.	69
5.1	Ordered eigenvalues of a shrinkage estimate, $\widehat{\Sigma}_\rho$, in comparison with the eigenvalues of a true covariance, Σ , and the sample covariance matrix, \mathbf{S} . Σ is of AR(1) structure with $b = 0.5$, and \mathbf{S} and $\widehat{\Sigma}_\rho$ are calculated from a sample of size $n = 25$ drawn from a multivariate ($p = 20$) normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix Σ	78
5.2	A hypothetical diagram illustrating the shrinkage effect for bivariate data (Warton, 2008). The lines are the 99% probability contours based on the true covariance (the solid lines) and shrunken covariance (the dashed lines). The shrinkage estimate reduces the eccentricity of the ellipse (the ellipse represented by solid black line is squashed along the major axis and make the ellipse represented by dashed line). The diagram also illustrates how the shift in different orientation can effect the power of a method to detect it. The red ellipse shows the shifted true distribution. It is shifted along the (a) first eigenvector (b) along second eigenvector. A larger shift is required along the first eigenvector to be detected as compared to the shift along second eigenvector. Note that “detected” means those red points that are outside the black ellipses.	79
5.3	Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for the Hotelling T^2 control chart (uses sample covariance matrix) and the blue lines are the results from new method.	84

5.4	Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for the Hotelling T^2 control chart (uses sample covariance matrix) and the blue lines are the results from new method.	85
5.5	First two principle components of the gene expression data. All out-of-control points in Figure 5.6 (including 12 out-of control points from baseline set of data) are encircled.	91
5.6	Multivariate control chart using study 1 data as a baseline set of data (black) and study 2 as a future set of data (red). The solid line at $T^2 = 853.504$ represent the UCL at 5% level of significance.	92
5.7	Principal components trajectory plot for the chemical process data with 99% confidence ellipse. Only the first 20 baseline observations are used to compute the principal components. The 10 future observation are plotted with star symbols and are numbered from 21 to 30 in order to show the natural sequence of the points.	93
5.8	Control chart produced by the proposed method for monitoring the chemical process data shown in Figure 5.7. The first 20 observations are used to estimate the control limits. The solid line at $T^2 = 11.1887$ indicates the control limit of the chart at 1% level of significance.	94
5.9	Principal components trajectory plot for the chemical process data with 99% confidence ellipse. Note that the first 10 baseline observations are dropped from the analysis and the principal components are computed from the middle 10 observations. The last 10 observations are plotted with star symbols and are numbered from 21 to 30 in order to show the natural sequence of the points.	95
5.10	Control chart produced by the proposed method for monitoring the chemical process data shown in Figure 5.9. Only the middle 10 observations are used to estimate the empirical reference distribution. The solid line at $T^2 = 13.1733$ indicates the control limit of the chart at at 1% level of significance.	96
A.1	Power comparison of MANOVA test based on 5 competing procedures under AR(1) covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the power is estimated using 1000 samples and the significance level is kept as 0.05.	107
A.2	Power comparison of MANOVA test based on 5 competing procedures under exchangeable covariance structure with $b = 0.4$. For each value of $p \in [2, 30]$, the power is estimated using 1000 samples and the significance level is kept as 0.05.	108

B.1	Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.3$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	110
B.2	Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.4$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	111
B.3	Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.6$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	112
B.4	Power (solid lines) and false alarm rate (dashed lines) for AR(1) covariance structure with $b = 0.7$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	113
B.5	Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.3$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	114
B.6	Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.4$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	115
B.7	Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.6$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method.	116

B.8 Power (solid lines) and false alarm rate (dashed lines) for exchangeable covariance structure with $b = 0.7$. The black lines are based on the true parameters and therefore are the best possible results one can achieve. The green lines are the results for standard method (using sample mean and sample covariance matrix) and the blue lines are the results from new method. 117

List of Tables

3.1	Mean squared errors of the estimated eigenvalues presented in Figure 3.6.	46
3.2	Mean squared errors of the estimated eigenvalues presented in Figure 3.7.	46
3.3	Mean squared errors of the estimated eigenvalues presented in Figure 3.8.	47
4.1	Values of the shift parameter, δ , used in the simulation study to obtain a moderate power.	62
4.2	Table of p-values for five competing procedures. Significant effects at $\alpha = 0.05$ are shown in bold.	68
5.1	In-control median run length (MRL) under both the AR(1) structure of covariance and the exchangeable structure of covariance with $b = .6$ (the lower and upper quartiles are shown inside the parentheses). The desired MRL is the median of the geometric distribution with parameter α . The size of the baseline set of data is 50.	88
5.2	In-control median run length (MRL) under both the AR(1) structure of covariance and the exchangeable structure of covariance with $b = .6$ (the lower and upper quartiles are shown inside the parentheses). The desired MRL is the median of the geometric distribution with parameter α . The size of the baseline set of data is 100.	89
5.3	Chemical process data. There are total 30 observations. The first 20 observations constitute the baseline set of data and the last 10 observations are the new observations used for testing and monitoring.	99
A.1	Values of the shift parameter, δ , used in the simulation experiments whose results are presented in Figure A.1 and Figure A.2.	106