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

List of Figures

List of Tables