SOME ASPECTS OF COVARIANCE REGULARISATION IN DISCRIMINANT ANALYSIS

A Thesis presented in fulfilment of the
requirements for the degree of
Doctor of Philosophy in Statistics at
Massey University, New Zealand.

John Peter Koolaard BSc.(Hons), MSc.

1997

Errata

to the thesis by J.P.Koolaard entitled "Some Aspects of Covariance Regularisation in Discriminant Analysis".

Page 1, line -10 'Prostrate' should be *prostate.*

Page 3, line 5 After the words "...for group k." add the sentence: "It is evident from expression (1.3) that all vectors in the thesis are considered as column vectors, unless stated otherwise."

Page 4, line 6 Sentence beginning on this line should read:"In effect, the $\mathbf{S}_k$ are replaced by the pooled covariance matrix, and the variance of the elements of $\mathbf{S}_p$ are smaller ..."

Page 15, line -11 "...where the pooled sample estimate..." should read: "...where the inverse pooled sample estimate...".

Page 56, lines -6 to -4 Rewrite these three lines as: "It should be noted that in his article, Friedman used robust covariance estimators in place of $\mathbf{S}_k$ and $\mathbf{S}_p$ in expressions (3.6) and (3.7). The resulting robustification of (3.6) is written as ..."

Page 76, lines -9, -8, -5 In these lines replace $\tilde{\mathbf{S}}_k$ with $\mathbf{S}_k$ and $\tilde{\mathbf{S}}_p$ with $\mathbf{S}_p$.

Page 3, line 14 Change "... expression (1.9) ..." to "... expressions (1.9) and (1.10) ...".

Page 5, line 12 To avoid any possible confusion, change "$(i, j = 1, \ldots, K(i \neq j))$" to "(for all $j(\neq i) = 1, 2, \ldots, K$)".

Page 5, line 15 Remove the word "directly".

Page 12, line 1 Change "mean" to "mean vector".

Page 14, line 8 Change "off diagonal" to "off-diagonal".

Page 17, line 10 Change "(1993)" to "(1993))".

Page 21, line 14 Change "samples of" to "samples of size".

Abstract

Statistical discriminant analysis and classification are multivariate techniques concerned with separating distinct set of objects, and with allocating new objects to previously defined populations or groups. In this process the covariance matrix plays an important role, and usually this matrix has to be estimated from sample data. In this thesis, attention is focussed on investigating the problem of (poor) estimation of the covariance structure and its effects in statistical discriminant analysis. The quality or statistical properties of these estimates usually affect the resultant classification rules which are constructed using them.

Reasons for the (usually, consistent) estimators of the covariance matrices being poor are mainly to do with the quality and/or size of the training sample in relation to the number of parameters which have to be estimated. In this thesis, we are interested in investigating this problem as it occurs in the small sample, high-dimensional situation. In particular, we are interested in the problem of covariance estimation in the situations when the sample size to dimension ratios are relatively small. The criterion used to determine the success or otherwise of various methods used to address this problem is the estimated (overall) error rate. One method of dealing with a situation which potentially results in poor estimation of the covariance matrix is to impose a prescribed (simple) structure on the covariance matrix, such as the identity matrix, or multiple of it. Another method is to make the assumption that all the groups have the same covariance matrix. The effect of such simplifying assumptions is to reduce the number of parameters to be estimated. Consequently, the (fewer) parameters are estimated with higher precision. It has been demonstrated that this may result in better statistical discriminant analysis, even if the simplifying assumptions may not be entirely correct.

Of the classification rules based on the normal distribution, the quadratic discriminant function (QDF) makes no restrictions on the population parameters, and as such is the most general of this class of classification rules. However, it is also the one most affected by poor population parameter estimates. The two common simplifying techniques mentioned earlier (i.e. imposing an identity matrix structure on the covariance matrix, or assuming a common covariance among all populations) lead to two other discriminant rules, namely, the Euclidean distance function (EDF, based on the Euclidean distance between the group means) and the *popular linear discriminant function (LDF, based on the Mahalanobis distance between the groups) respectively. The sample-based versions of these two classifiers are compared using expected error rates (conditional on a set of training data), and these expected error rates are obtained through the derivation of asymptotic expansions. The expansions are evaluated under a range of settings, defined by employing combinations of various values of dimension, group separation, and covariance structure. It is shown that the simpler sample Euclidean distance function (SEDF) performs as well as or better than the sample linear discriminant function (SLDF) under most of the settings used. Exceptions occurred when the Mahalanobis distance between populations was much greater than the Euclidean distance.*

A flexible discrimination model, or rather, class of models, was developed by Friedman (1989), and called the regularised discriminant function (RDF). The sample version of the RDF (i.e. SRDF) model incorporates the general sample quadratic discriminant function (SQDF), the two previously-mentioned restricted models (SEDF and SLDF), as well as a wide range of models intermediate to these, through the use of additional "regularisation" parameters. The method employs two types of shrinkage of the covariance estimates - towards the pooled estimate on one hand, and towards a multiple of the identity matrix on the other. A separate regularisation parameter controls shrinkage to each. The training data is used in the model selection process to determine appropriate values for the regularisation parameters, through the use of cross-validation. The quality of model selection procedure which specifies a discriminant model is a crucial factor, since if it is performing well, it will result in a classification rule close to the optimal one from the class of models available.

Through large-scale simulation studies, the performance of the sample regularised discriminant function (SRDF) is investigated and it is shown that the SRDF generally leads to lower overall error rates than the standard classification rules. This is found to be largely due to the facility which allows shrinkage of the covariance matrices to sphericity, or eigenvalue regularisation. It is also found that the SEDF performs very well in relation to the SRDF for a variety of settings. Further simulation studies show that the performance of the SRDF is more sensitive to the parameter controlling shrinkage to sphericity than the one controlling covariance mixing. Also, it is found that under some circumstances, the SRDF performs better than the other classifiers even for quite large sample size to dimension ratios.

A crucial negative feature of the SRDF is its lack of scale invariance. The cause of this is eigenvalue regularisation. A modified classification rule is developed which is scale invariant, and is compared to the SRDF and the other classifiers via simulation. The modified rule omits eigenvalue regularisation, but otherwise increases sensitivity to the data by allowing for varying degrees of shrinkage to the pooled covariance for each group. It is shown that eigenvalue regularisation is generally beneficial for discrimination in medium to large dimensional problems, through its variance-reduction effect which stabilises the covariance estimates. Thus, the study concludes that scale invariance must be sacrificed in order to achieve reductions in error rate, in the absence of a suitable replacement for eigenvalue regularisation.

The use of cross-validation in the model selection process of the SRDF is also investigated, for several reasons: the computational effort involved, and the fact that it rarely leads to a unique choice of model, and often uses only a small subset of the available observations, in the model selection process. Consequently, another method for determining the optimal regularisation parameters is investigated. In particular, it is investigated whether appropriate values for the regularisation parameters can be indicated from a measure of the distance between the groups. For this purpose, the Bhattacharyya distance is chosen since it comprises a term primarily pertaining to the difference between group means, and a further term which indicates the level of disparity between group covariance structures. It is shown that the magnitudes of the various components of the Bhattacharyya distance, when considered on their own and in relation to each other, do give information as to appropriate values for the regularisation parameters. A new simulation study, as well as various case studies are presented to assess the performance of a new regularised discriminant function which uses the Bhattacharyya distance estimates between groups to select regularisation parameters for given training data. This classifier is shown to perform as well as the SRDF, and is computationally much faster since it avoids any re-sampling methods.

It is clear that most of the investigations and assessments of the various regularised discriminant rules have to be undertaken using Monte-Carlo simulation techniques, especially to estimate error rates. This is because exact analytical expressions for the unconditional error rate of the SRDF do not exist, except in certain limited circumstances. It has not been possible to obtain asymptotic expansions or some form of approximations of these error rates in a general context.

However, an approximation which can be used to calculate algebraically the error rate of the SQDF, assuming known population parameters under (other) strict conditions, is available in the literature. This approximation is used in this thesis to further examine the effects (observed in earlier simulation work) of the covariance regularisation parameters on error rates. This is the last piece of work in the thesis and, in spite of its limited extent (because of the restricted conditions of the approximations given), it largely confirms the results which were obtained from simulation experiments in the previous parts of the thesis.

Acknowledgements

I would like to thank my chief supervisor, Charles Lawoko, now at the Queensland University of Technology, for his immense help and guidance throughout this project. His office always had an 'open' door for me, and his patience with this part-time student was considerable. Likewise, I would like to thank my second supervisor, Ganesalingam, for his frequent doses of moral support. His too was an open door for me, and I was privileged to have supervisors such as these to sit with and think about the problems at hand.

The staff in the Statistics Department have treated me wonderfully, and I'd like to especially thank Professor Jeffrey Hunter, former Head of Department for his confidence in me, and his general concern and kindness towards me and my family, expressed in a number of tangible ways throughout my time in the Department.

Thanks also to Professor Dick Brook, an old golfing partner, who always kept an eye out for me, and to whom I could always go to for a chat. To Paula, Helen, Greg, Chin Diew, Ganesh and the others too, thanks.

The support of Massey University for this project must also be acknowledged. I am grateful for the support of a Massey University Doctoral Scholarship for 1995, as well as for part-time employment in the the form of Assistant Lectureships for four years from 1991 to 1994.

It has been a slow, long, hard grind all the way, and one who knows that more than all the others is my dear wife Antoinette. Hers was a constant support, including necessary admonishment at times. She had to put up with a lot from me, and I'm very grateful for it all. It was and is always a joy to come home to Antoinette and my two lovely daughters, Lydia and Natasha.

Finally, thanks and praise be to the Father of my Lord Jesus Christ, who carried me through it all, and who sustains and upholds me.

List of Additional Publications by Author Including Papers Presented at Conferences
See Appendix C

1. **Koolaard**, J. P. and Lawoko, C. R. O. (1993). Estimating error rates in discriminant analysis with correlated training observations: a simulation study. *J. Statist. Comput. Simul.* 48, 81-99.

2. **Koolaard**, J. P. and Lawoko, C. R. O. (1994). Some results on the error rates of the Euclidean and linear discriminant functions. *Proceedings of the ORSNZ/NZSA Conference, Massey University, Palmerston North, New Zealand* (August 1994). pp 327-332.

3. **Koolaard**, J. P. (1995). Covariance Shrinkage in Discriminant Analysis. Paper presented to the A. C. Aitken Centenary Conference, Dunedin, New Zealand (August 1995). [Winner of SPSS Statistics Prize for best statistics paper presented by a student.]

4. **Koolaard**, J. P., Lawoko, C. R. O. and Ganesalingam, S. (1996). Regularized discriminant (classification) analysis involving Bhattacharya distance measure. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 2, Poster, pp 35-43.

5. **Lawoko**, C. R. O., and Koolaard, J. P. (1996). Applications of regularised discriminant(classification) functions in the classification of objects: a discussion of potential applications to remote sensing. *Proceedings of the 8th Australasian Remote Sensing Conference, Canberra, Australia* (March 1996). Volume 1, pp 177-184.

6. **Koolaard**, J. P., Ganesalingam, S. and Lawoko, C. R. O. (1996). Comparison of regularised discriminant analysis with the standard discrimination methods. Paper presented to the International Biometrics Conference (IBC '96), Amsterdam, the Netherlands (July 1996). Also submitted to the Journal of Classification.

7. **Koolaard**, J. P. and Lawoko, C. R. O. (1996). The linear and Euclidean discriminant functions: a comparison via asymptotic expansions and simulation study. *Commun. Statist.- Theory Meth.*, (To appear).

Contents

List of Figures

List of Tables

List of Abbreviations used in this Thesis

SLDF	Sample linear discriminant function.
SQDF	Sample quadratic discriminant function.
SRDF	Sample regularised discriminant function, similar to the method developed by Friedman (1989).
SRDF1	Rule based on SRDF, but where a policy of minimum regularisation (instead of maximum regularisation as with SRDF) is employed to break ties in cases where the model selection procedure does not yield a unique choice of values for the regularisation parameters.
SRDF-M	A modified regularised rule which omits the eigenvalue shrinkage parameter γ but allows for as many covariance mixing parameters (λ's) as there are groups to be discriminated between. This rule is scale-invariant, unlike the SRDF.
SRDF-M1	Similar to SRDF-M, but where a policy of minimum regularisation is employed to break ties (as for SRDF1).
SRDF-B	Regularised discriminant rule which chooses the λ and γ parameters by using information obtained from a measure of the Bhattacharyya distance between pairs of populations (of interest).
SEDF	Sample Euclidean distance function. In this thesis, this rule is formed by setting the regularisation parameters (λ and γ) in the SRDF rule both equal to one.