# High Breakdown Linear Discriminant Analysis

Douglas M. Hawkins and Geoffrey J. McLachlan*

University of Minnesota and University of Queensland

September 4, 1998

**Abstract**

The classification rules of linear discriminant analysis are defined by the true mean vectors and the common covariance matrix of the populations from which the data come. As these true parameters are in general unknown, they are commonly estimated by the sample mean vector and covariance matrix of the data in a training sample randomly drawn from each population. These sample statistics are however notoriously susceptible to contamination by outliers, a problem compounded by the fact that the outliers may be invisible to conventional diagnostics. High breakdown estimation is a procedure designed to remove this cause for concern by producing estimates that are immune to serious distortion by a minority of outliers, regardless of their severity. In this paper, we motivate and develop a high breakdown criterion for linear discriminant analysis and give an algorithm for

*Douglas M. Hawkins is Professor, Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108, U S A. Geoffrey J. McLachlan is Reader, Department of Mathematics, University of Queensland, Brisbane, Australia. This work was carried out while the first author was Raybould Visiting Scholar at the University of Queensland. The work of the second author was supported by the Australian Research Council.

its implementation. The procedure is intended to supplement rather than replace the usual sample-moment methodology of discriminant analysis either by providing indications that the data set is not seriously affected by outliers (supporting the usual analysis) or by identifying apparently aberrant points and giving resistant estimators that are not affected by them.

**KEY WORDS:** Minimum covariance determinant, Outliers, Classification rules.

# 1 General Background

Normal-theory-based discriminant analysis is very widely used for the classification of multivariate data into different possible source populations. The underlying model is that the $p$-dimensional vectors of training data $\boldsymbol{Y}_{ij}$ follow the distribution

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (i = 1, 2, ..., g;\ j = 1, ..., n_i),$$

where $n_i$ represents the size of the training sample from source population $i$, for each of the $g$ different populations, or groups. This framework specializes into two different problems. The heteroscedastic discriminant analysis problem (leading to the 'quadratic discriminant function') arises when the covariance matrices $\boldsymbol{\Sigma}_i$ are unrestricted. The more specialized homoscedastic discriminant analysis model (leading to the 'linear discriminant function') arises when the $\boldsymbol{\Sigma}_i$ are all equal to a common covariance matrix $\boldsymbol{\Sigma}$.

The parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ or $\boldsymbol{\Sigma}$ provide the means to classify future unknowns which are believed to come from the same source populations. The Bayes rule for optimal classification of future unknowns is defined by these parameters, together with the prior probabilities with which the unknown comes from each of these possible sources. The minimum-cost classification rule is obtained by adding further information on the costs associated with various possible misclassifications. McLachlan (1992) provides a thorough coverage of these and other issues associated with the discriminant analysis model.

If the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ were known, then we would have no need for the training data $\boldsymbol{Y}_{ij}$ except possibly for the information they provide on the prior probabilities of a future unknown coming from each of the sources. In practice though the parameters are not known, and it is the function of the training data to help provide estimates of them. The standard estimates are the maximum likelihood estimates (MLE's) or their unbiased variants. The sample mean vectors for the provision of the estimates of $\boldsymbol{\mu}_i$ are given by

$$\hat{\boldsymbol{\mu}}_i = \bar{\boldsymbol{Y}}_{i.} = \sum_{j=1}^{n_i} \boldsymbol{Y}_{ij}/n_i \qquad (i = 1, ..., g).$$

For the heteroscedastic case, each $\boldsymbol{\Sigma}_i$ is estimated by its training sample analogue, usually after correction for bias, to give

$$\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i = \sum_{j=1}^{n_i} (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.})'/(n_i - 1) \quad (i = 1, ..., g).$$

For the homoscedastic discriminant analysis model the standard estimate of the common covariance matrix $\boldsymbol{\Sigma}$ is the pooled within-group sample covariance matrix (corrected for bias)

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{Y}}_{i.})'/(n - g),$$

where n $= \sum n_i$ is the total sample size across all groups.

These simple, easily-computed MLE's provide the starting point for most derivations of classification rules. Their simplicity, however, may conceal the fact that they are extremely susceptible to outliers. A single outlier in the overall data set can be enough to distort not only the sample mean vector of the group from which it came, but also the pooled sampled covariance matrix, and can lead to an arbitrarily poor classification rule.

This observation has led to the suggestion of using more robust methods to estimate the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. The heteroscedastic case is immediately amenable to conventional single-sample robust estimation, since its estimation decomposes into the separate estimation of the mean vector and covariance matrix of each of the source populations, with the implication that any method for estimating the mean vector and covariance matrix of a single multivariate sample may be applied directly and without special adaptation to the estimation of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Early robust estimation proposals concentrated on $M$-estimation. The general framework of $M$-estimation is covered in considerable detail in the books by Huber (1981), and Hampel, Ronchetti, Rousseeuw and Stahel (1986), while McLachlan (1992) summarizes the use of $M$-estimation in discriminant analysis. In view of this existing thorough coverage, we will not expand on this topic.

The weakness of $M$-estimation is its low breakdown point of $1/p$. The breakdown point is the minimum proportion of badly-placed outliers that can drive the estimators arbitrarily far from their true values. The breakdown point of $1/p$ of $M$-estimators means that in 20 dimensional data, for example, $M$-estimation cannot be relied on in data sets that could have more than 5% of outlying cases. This low level of protection against outliers is frequently unacceptable.

Instead of $M$-estimates, one may use such high breakdown estimates as the minimum volume ellipsoid (MVE) or the minimum covariance determinant (MCD) estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$; see, for example, Rousseeuw and Leroy (1987), and Hampel et al. (1986). These provide the assurance of protection against much higher proportions of bad cases. Other high breakdown possibilities include $\tau$- and $S$-estimates which provide higher statistical efficiency at the cost of greater computational load (see, for example, Tyler (1991) and Lopuhaä (1991) for a general discussion of these approaches to high breakdown multivariate estimation). In the heteroscedastic case, this consists of simply applying standard single-sample high breakdown estimation methods to each of the groups separately.

The homoscedastic case does not yet have a standard high breakdown methodology. It is possible to apply single-sample estimation, as is done for example in Chork and Rousseeuw (1992). These authors use conventional high breakdown estimates based on the MVE for each group separately, and then pool the resulting group-specific covariance matrices to obtain a pooled covariance matrix. While providing a high breakdown point, though, this approach has some deficiencies. It restricts the number of outlying cases that could occur in each of the groups separately, and by applying equal trimming to all groups, leads to a potential loss of statistical efficiency in situations where some groups are outlier-free. For this reason, we will now derive a high breakdown methodology specially adapted to the homoscedastic discriminant analysis setup, and show its connection to existing high breakdown technologies for the single-sample case.

## 2 The Mean-Shift Outlier Model

To help motivate the methodology, we broaden the homoscedastic discriminant analysis model to accommodate the possibility of outliers. There are several ways of doing this, but one convenient general approach is that of the mean-shift outlier model. Recast the defining model as

$$\boldsymbol{Y}_{ij} \sim N(\boldsymbol{\mu}_i + \boldsymbol{\alpha}_{ij}, \boldsymbol{\Sigma}) \quad (i = 1, 2, ..., g;\ j = 1, ..., n_i), \tag{1}$$

where the new parameters $\boldsymbol{\alpha}_{ij}$ are mean-shift parameters. The outliers are characterized by having arbitrary nonzero mean shifts, while the inliers have $\boldsymbol{\alpha}_{ij} = \mathbf{0}$. Both identifiability and realism require that only a minority of the cases be outliers. The identity of the outlying cases is not known *a priori*, and it is necessary to simultaneously identify the cases most likely to be outliers and to estimate the parameters. Write $C$ for the number of inliers, so that by implication $n - C$ is the number of outliers. We will assume that $C$ is known; this is not a serious restriction to the applicability of the method –since an outlier with $\boldsymbol{\alpha}_{ij} = 0$ is an inlier, all that is required for $n - C$ is an upper limit on the number of outliers that might be present in the sample. Specifying some reasonable upper limit for $n - C$ will give the flexibility to handle any number of outliers up to $n - C$.

The objectives of identifying the most likely outliers and estimating the parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}$, and (if of interest) $\boldsymbol{\alpha}_{ij}$ may be met using maximum likelihood. The doubled log likelihood for the model is

$$2\log L = -\sum_{i=1}^{g}[n_i \log|\boldsymbol{\Sigma}| - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij} - \boldsymbol{\alpha}_{ij} - \boldsymbol{\mu}_i)(\boldsymbol{Y}_{ij} - \boldsymbol{\alpha}_{ij} - \boldsymbol{\mu}_i)')],$$

where the $\boldsymbol{\alpha}_{ij}$ term is only present for the outlying terms. Standard likelihood maximization arguments then give the estimates

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_i &= \sum_{j=1}^{n_i}\delta_{ij}\boldsymbol{Y}_{ij}/\sum_{j=1}^{n_i}\delta_{ij},\\
\hat{\boldsymbol{\alpha}}_{ij} &= (1-\delta_{ij})(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_i),\\
\hat{\boldsymbol{\Sigma}} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\delta_{ij}(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_i)'/n,
\end{aligned}$$

where $\delta_{ij}$ is an outlier indicator whose value is 0 for each case that is modeled as outlying and 1 for each case that is modeled as inlying. The maximization involves two stages - the decision of which $C$ of the cases are to be assigned the zero $\delta_{ij}$ values, (we call these the "covered" cases) followed by the trivial step of computing the MLE's for this choice of indicators. The cases that are assigned $\delta_{ij} = 1$ we will describe as "uncovered"; they are identified as possible outliers by the maximum likelihood criterion.

The maximized log likelihood is a linear function of the fitted generalized variance $\left|\hat{\boldsymbol{\Sigma}}\right|$ with a negative coefficient. This shows that the cases with the greatest likelihood of being the outliers are those cases whose deletion leads to the smallest value of the generalized variance (that is, the determinant of the pooled within-group covariance matrix of the covered cases).

We will abbreviate this criterion to the MWCD, for Minimum Within-Group Covariance Determinant criterion, and use it as the basis for our proposal for high breakdown discriminant analysis.

It should be noted that in the case of a single group, our MWCD criterion reduces to the minimum covariance determinant (MCD) criterion which is in widespread use as a high breakdown estimator for the case of a single random sample (see, for example, Rousseeuw and Leroy (1986)), and can be motivated by an analogous likelihood argument. The MWCD criterion is thus a generalization of the MCD criterion to a homoscedastic multigroup setting.

## 2.1 The breakdown point of the MWCD

The coverage $C$ defines the maximum number of outliers that can be accommodated. We will now look more closely at the limitations that there are on the possible values of $C$, and by extension at what the breakdown point of the MWCD estimators is. We will do this informally for the sake of readability; a formal derivation of the breakdown point of the MWCD closely follows that for its single-group specialization, the MCD estimator, discussed in Rousseeuw and Leroy (1987) (Theorem 1 page 258, and pages 262 and 264).

The breakdown point of an estimator is defined in terms of a 'game against nature' based on the number of observations in the original data set that must be changed in some suitable way for the resulting change in the estimator to be unbounded. In the multigroup location/scale problem, there are two types of such complete failure in the estimator - 'implosion', resulting in $\hat{\boldsymbol{\Sigma}}$ becoming a zero matrix; and 'explosion' in which $\hat{\boldsymbol{\mu}}$ becomes unbounded.

We will make the assumption that the data are in 'general position', a standard high breakdown assumption, which in the context of the multiple group discriminant analysis problem means that the within-group pooled covariance matrix obtained from any subset of the data vectors of total size $p + g$ will be non-singular. If the data vectors really do follow a continuous distribution, this will occur with probability 1.

Suppose the MWCD approach results in the retention of $m_i$ observations from group $i$ (the cases whose $\delta_{ij} = 1$), with the remaining $n_i - m_i$ cases (those with $\delta_{ij} = 0$) being trimmed. The total coverage is $\sum_1^g m_i = C$ by definition. Suppose for ease of notation that the covered cases are the first $m_i$ in each group, and the trimmed cases the subsequent ones. Then the MWCD estimators can be written

$$\begin{aligned}\hat{\boldsymbol{\mu}}_i &= \sum_{j=1}^{m_i} \boldsymbol{Y}_{ij}/m_i, \\ \hat{\boldsymbol{\Sigma}} &= \sum_{i=1}^{g}\sum_{j=1}^{m_i}(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_{ij} - \hat{\boldsymbol{\mu}}_i)'/n.\end{aligned}$$

By further routine orthogonal transformations, we may re-express the covariance matrix estimate as

$$\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{g}\sum_{j=2}^{m_i} \boldsymbol{Z}_{ij}\boldsymbol{Z}'_{ij}/n.$$

Now consider perturbations of the data that could lead to implosion. Start with any subset of size $p + g$ of the covered cases, taking at least one from each group. This subset will necessarily give a singular pooled within-group covariance matrix because of its rank deficiency. Now replace the uncovered data vectors in group $i$ with arbitrary linear combinations of the subsampled $\boldsymbol{Z}_{ij}$ in their group. If we replace enough uncovered data vectors with such linear combinations, and swap them for the covered cases not in the subsample, the resulting $\hat{\boldsymbol{\Sigma}}$ will be singular, leading to the method's implosion.

For this breakdown to be possible requires a sufficiently large number of uncovered cases $n - C$ to be available for modification and to replace the $C - (p + g)$ covered but not sub-sampled cases - specifically

$$n - C \geq C - (p + g)$$

or

$$C \leq \frac{n+p+g}{2}$$

Guaranteed avoidance of implosion therefore implies that C must exceed this lower bound, giving the requirement

$$C > \frac{n+p+g}{2}$$

.

Turning next to the mean vectors, the covered cases necessarily have the smallest Mahalanobis distances from their sample mean. The estimate $\hat{\boldsymbol{\mu}}_i$ could therefore be made to explode by replacing a sufficiently large number of uncovered cases by some arbitrarily large constant vector. To be able to do this would require that there be at least as many uncovered as covered cases:

$$n_i - m_i \geq m_i$$

and avoiding this potential for explosive breakdown requires the converse inequality

$$m_i > \frac{n_i}{2}.$$

A guarantee of no breakdown therefore requires that the coverage be more than half the cases in each of the groups, and a total of more than $(n+p+g)/2$.

## 2.2 Removing bias from the estimators

One concern when using robust methodologies is that they should agree asymptotically with the standard nonrobust methodologies in large samples which do not contain any outliers. In the context of discriminant analysis, this invites the question of the statistical consistency of the estimators when applied to normal data. Symmetry considerations show that the MWCD criterion leads to unbiased estimates of the mean vectors $\boldsymbol{\mu}_i$, but the trimming leads to $\hat{\boldsymbol{\Sigma}}$ being a biased estimate of the true covariance matrix $\boldsymbol{\Sigma}$ - its expectation is some scalar multiple of $\boldsymbol{\Sigma}$.

Since the primary objective of discriminant analysis is to provide good classification rules rather than to estimate parameters, this bias in $\hat{\boldsymbol{\Sigma}}$ is not necessarily of great concern. It matters only to the extent that it impacts on the classification rule. For discriminant analysis with equal prior probabilities for the groups, all that is required of $\hat{\boldsymbol{\Sigma}}$ is that it be consistent for some multiple of $\boldsymbol{\Sigma}$; the constant of proportionality does not affect the subsequent classification rules at all. However, the bias in $\hat{\boldsymbol{\Sigma}}$ does affect the classification rules when prior probabilities or misclassification costs are unequal, and it also affects the posterior probabilities of membership of the groups in the equal-prior case. For these reasons, finding a modification of $\hat{\boldsymbol{\Sigma}}$ that is consistent for $\boldsymbol{\Sigma}$, or even better, unbiased for $\boldsymbol{\Sigma}$ merits some attention even though it is not central to the success of the method.

The issue of consistency is resolved by Butler, Davies and Jhun (1993) for the MCD criterion. Fixing $p$, setting $C = \lambda n$ for a fixed coverage proportion $\lambda$, and letting $n \to \infty$, they show that the MCD estimator $\hat{\boldsymbol{\Sigma}}$ converges weakly to $\rho(\lambda)\boldsymbol{\Sigma}$. The constant of proportionality $\rho(\lambda)$ is given by

$$\rho(\lambda) = F(k|p+2)/F(k|p). \tag{2}$$

where the notation $F(x|m)$ represents the cumulative distribution function of a chi-squared distribution with $m$ degrees of freedom, and $F(k|p) = \lambda$; that is, $k$ is the quantile of order $\lambda$ of the chi-squared distribution with $p$ degrees of freedom. Table 1 shows the value of the constant $\rho(\lambda)$ for various values of $p$ and $\lambda$.

In the multigroup situation of concern here, the development of Butler et al. can be adapted to show that the same limit holds provided $n \to \infty$, with $g$ and $p$ held constant and $C = n\lambda$. This means that dividing $\hat{\mathbf{\Sigma}}$ by the constant $\rho(\lambda)$ (which depends only on $p$ and $\lambda$) will yield an estimator consistent for $\mathbf{\Sigma}$.

The removal of bias in small samples is less amenable to theoretical calculations. Some limited simulations have suggested that even in samples of realistic but small size, dividing the MWCD covariance matrix estimate by $\rho(\lambda)$ provides estimates with little bias and that are more than adequate for the comparative calculations involved in discriminant analysis.

Also shown in Table 1 is the asymptotic efficiency of the mean vector as estimated by the MCD. These efficiencies bear on the relative attractiveness of our proposal and that by Chork and Rousseeuw of using separate high breakdown estimates of the mean vector and covariance matrix in each group and then combining the separate estimates. Chork and Rousseeuw proposed estimating the group mean vectors and covariance matrices using the minimum volume ellipsoid. As this estimator has an efficiency of zero, we refine their proposal by substituting the MCD. To allow for possible maldistribution of the outliers across the groups then, it is essential in their proposal to use a small coverage within each group – following Chork and Rousseeuw's lead therefore we might select 55% coverage. For say $p = 4$, this would then give an efficiency of 28% for the estimates of $\boldsymbol{\mu}_i$. Our proposal would gain by being able to use a higher overall coverage and still have the ability to handle individual groups with large numbers of outliers. If for example we select 80% coverage the efficiency of the estimates of $\boldsymbol{\mu}_i$ would be 58%. This greater efficiency in locating the true means will have a corresponding impact on the quality of the classifications made by the discriminant analysis. Increased coverage will also lead to improved estimates of the covariance matrix, providing an additional source of improvement over the approach of separate high breakdown estimates for each population. This then provides support for the notion of concentrating the trimming where it seems to be needed (as in our proposal) rather than applying trimming across the board.

## 2.3 The choice of $C$

A final comment is on the choice of the coverage $C$. This may be approached from at least four directions. Notions of statistical efficiency argue for making $C$ as large as possible so that the estimates of the parameters have as much statistical efficiency as possible consistent with the possible presence of outliers. Standard practice in high breakdown estimation is in the opposite direction – to make $C$ as small as possible (but not below the lower limit required to guarantee that breakdown is impossible) so as to maximize the number of outliers $n - C$ that can be accommodated. This is usually not necessary in the context of discriminant analysis, since the objective is to provide protection against the maximum number of outliers we think it likely to find in data, rather than the maximum number that could theoretically be accommodated. Thus in many application contexts there may be reason to believe that the number of outliers would not be likely to exceed some specified maximum, (say for example 10% or 25% of the data), so a third approach is then for $C$ to be selected to provide for this maximum plausible number of outliers: $0.9n$ in the first case, and $0.75n$ in the second. The fourth approach, suggested by Cook and Hawkins (1990) is to evaluate the solutions for a range of coverages, comparing the solutions for the different coverages. This is perhaps the best of the three possibilities, though the one requiring the most computation.

## 3 An Algorithm for Finding the MWCD

The MWCD criterion involves the identification of the subset of cases whose deletion leads to the smallest value of the determinant of the within-group covariance matrix. It is well established that this criterion function can not be found reliably by simple methods such as sequential deletion of the cases with largest Mahalanobis distance from their group means. The closely-related single-sample MCD criterion has been studied more extensively, and has been shown to be handled very effectively by the 'feasible solution' algorithm of Hawkins (1994). This algorithm proceeds by starting with a random selection of the $C$ cases labeled as inliers, and then seeks pairwise swaps of a currently inlying case to replace a currently outlying case, looking for swaps that reduce the determinant of the covariance matrix. When no such pairwise swap can be found, the current labeling is termed a 'feasible solution', since it satisfies a necessary condition for the optimum. The global optimum is one such feasible solution, but not necessarily the only one. The full feasible solution algorithm operates by repeating this iteration from many different random starts, taking the best solution found to be the true optimum.

The MWCD criterion has a similar necessary condition – that its value can not be improved by any pairwise swap of a case currently labeled as an inlier for one currently labeled as an outlier – and this necessary condition leads to a feasible solution algorithm for the MWCD criterion also. In it, the user specifies the required coverage $C$. Then $C$ of the cases are selected at random and provisionally labeled as inliers. A search is made for pairwise swaps of a current inlier for a current outlier that lead to a reduction in the determinant of the within-group covariance matrix. If swaps can be found that lead to such an improvement, the swap giving the maximum improvement is made and a fresh attempt made to find another pairwise swap. When no swap leads to an improvement, then a feasible solution has been reached. The procedure is then repeated a number of times with a fresh random selection of $C$ provisional inliers, leading to a (possibly different) feasible solution. If there is more than one feasible solution, that with the smallest within-group covariance determinant is used as the solution.

A feasible solution procedure implementing this method has been written, and was used for the calculations of the examples discussed below. Concerns that a potential user might have relate to its execution time, and the likelihood that it reaches the globally optimal MWCD. In any particular data set, a count of the number of feasible solutions found provides the user with some, albeit unquantified, indication of whether the global optimum is likely to have been found. If for example, 1000 random starts are able to locate only 4 different feasible solutions, then it is entirely plausible that the best of these is indeed the global optimum. If on the other hand 100 random starts were to produce 70 distinct feasible solutions, the best of these could at most be described as a good local optimum. Such a good local optimum might still provide excellent classification rules; the major objection that one would raise to it is that not being the global optimum, it might not be replicated by a different analyst using a different random number stream.

Attempts to put this informal reasoning on a more formal basis (using for example capture-recapture ideas) are complicated by the fact that good local optima tend to be reached from more starting points than poor local optima, so that assumptions about the exchangeability of the local optima are not well founded. For this reason, we have not pursued the attempt to turn the feasible solution discovery history into a probability that one has found the global optimum.

Our computational experience has been that there are few feasible solutions (fewer than 10) in data sets of the size typically discussed in the discriminant analysis literature, though the number of feasible solutions can be expected to increase with the sample size.

As an indication of both the execution time and number of feasible solutions one might anticipate, running the familiar Fisher *Iris* data with 80% coverage on a 100MHz Pentium computer took 16 minutes to process 1,000 random starts. These produced 3 feasible solutions, the best of which happened to be found at the second random start. The paucity of feasible solutions gives good reason to think that this is the global optimum, and that the use of as many as 1,000 random starts gave substantial overkill for a modest computational cost.

# 4 Examples

## 4.1 The Diabetes Data

We start by illustrating the approach using the data set analyzed by Reaven and Miller (1979) in their investigation of the relationship between chemical diabetes and overt diabetes in 145 nonobese adult subjects. Attention in their study was focussed on the three variables (1) area under the plasma glucose curve for the three-hour oral glucose tolerance test (OGTT), (2) area under the plasma insulin curve for the OGTT, and (3) steady state plasma glucose response (SSPG). Using conventional clinical criteria, Reaven and Miller (1979) initially classified the 145 subjects on the basis of their plasma glucose levels into three groups, $G_1$ (normal subjects), $G_2$ (chemical diabetes), and $G_3$ (overt diabetes). The sample mean vectors and pooled within-group sample covariance matrix of the data are given in Table 2.

We ran this data set using the feasible solution algorithm for the MWCD with a range of coverages – from 100% (that is using a conventional homoscedastic discriminant analysis with no trimming at all) down to 60% in steps of 5%. At each coverage, 1000 random starts were made. The common log of the minimized within-group covariance matrix determinant (the generalized variance) at the different coverages is displayed in Table 3.

Visually, the determinant appears to drop quite rapidly going down to 80% coverage and more slowly after that, suggesting descriptively at least that one might wish to accommodate up to 20% outliers in this data set, but that little is gained by accommodating more. Note that the log determinant at say 75% coverage is smaller than that of the full data set by 3.84. This means that the covariance matrix following trimming is some 47 times smaller than that of the full data set. This gives some feeling for the degree of aberration of the cases that were trimmed.

The purpose of the discriminant analysis is derivation of good rules for reclassifying future unknowns, and it is against this criterion that the high breakdown MWCD should be assessed. We therefore tested the classical nonrobust and the MWCD procedures on the training data, using the leave-one-out cross-validation method.

The results of this comparison are displayed in Table 4. At 95% coverage, the robust and nonrobust analyses produce identical cross-validated reclassifications, but as the level of trimming increases, the high breakdown analysis starts to produce significantly fewer misclassifications than does the conventional nonrobust analysis. At high levels of trimming, one would expect to see the high breakdown method start to deteriorate as a consequence of its decreasing statistical efficiency, but deterioration is not really apparent over this range of coverages. The superiority of the high breakdown MWCD is most marked at 85% coverage.

Table 5 shows the mean vectors and bias-corrected covariance matrix for the solution with 85% coverage. It is striking that the mean vectors are similar to those of the full-sample case, while the variances are substantially smaller. This paints a picture of the trimmed cases being spread widely, but not in any particular preferred direction, from the group means. It is also noteworthy that most of the cases (15) were trimmed from group 3 and the remaining five from group 2, with group 1 left intact.

Another useful benchmark for comparison is Chork and Rousseeuw's (1992) approach of finding the MVE in each of the groups separately and pooling the estimates. We applied this method, using the 'feasible solution' algorithm for the MVE as set out in Hawkins (1993) to identify the most inlying cases within each group. These were then used to train the homoscedastic discriminant analysis. Somewhat surprisingly, this gave almost identical results to those of the conventional zero-breakdown discriminant analysis. The two methods differed on only two cases, with the Chork-Rousseeuw method better on one and worse on the other. The Chork-Rousseeuw approach therefore seems markedly inferior to the MWCD on this data set.

As noted earlier, using a probabilistic search method like the feasible solution algorithm raises question of how plausible it is that the solution returned by the search is the global optimum. It is therefore of interest to know how many different feasible solutions turned up in the analysis, and what fraction of the starting points led to the best solution found. These figures are reported in Table 6. None of the coverages had more than a handful of feasible solutions, and this shows that our use of 1,000 random starts was far in excess of what was needed to uncover the optimum reliably – even in the worst case one or two dozen would have given the same solution with high probability.

## 4.2 The Hemophilia Data

A second example is the Hemophilia data set which we report in much less detail. It was analyzed originally by Habbema, Hermans and Van den Broek (1974) in the context of discriminating between normal women and hemophilia A carriers on the basis of two feature variables, $x_1 = \log_{10}(\text{AHF activity})$ and $x_2 = \log_{10}(\text{AHF-like antigen})$. We let $G_1$ be the group corresponding to the noncarriers or normals and $G_2$ the group corresponding to the carriers. The classified training data consist of $n_1 = 30$ observations on known noncarriers and $n_2 = 45$ observations on known obligatory carriers. We applied the analysis with a single coverage of 75%. Table 7 shows the mean vectors and covariance matrix of the full data set, and Table 8 that for the subset covered by the MWCD. In 1000 random starts, three feasible solutions were found, with the solution shown in Table 8 being returned in 846, or 85% of the starts, again giving a high degree of confidence that this is the global optimum.

The common logarithms of the two determinants are –8.2905 and –9.5002 respectively, showing that the full sample within-group covariance matrix has a determinant only $e^{1.21} = 3.4$ times larger than that of the most compact 75% of the data, compared to a factor of 47 for the diabetes data. This more modest-looking difference suggests that there may not be much gain from using the high breakdown estimator, and might even be some loss because of its lower statistical efficiency.

Applying the leave-one-out cross-validation test to this data as analyzed by both procedures gives the summary counts shown in Table 9. Here there is very little to choose between the two approaches. In particular, while the high breakdown method did not do appreciably better than the nonrobust approach, neither is there any reason to think it did worse. While this may be advanced to argue that there is no need to use the high breakdown MWCD, it can equally well be used to argue the opposite – that using the MWCD provided protection against the possibility of there being outliers in the data, and that this protection came at no appreciable cost in terms of performance.

## 5 Conclusion

Discriminant analysis can be a very effective tool in multivariate multigroup data analysis, but suffers from its susceptibility to outlying cases. The traditional way of trying to circumvent this sensitivity has been to compute model fit diagnostics (such as, for example, those of Hawkins (1981)). This method is unfortunately not necessarily reliable, as it is possible to have arbitrarily severe multivariate outliers with quite innocuous full-sample diagnostics.

We have proposed a high breakdown approach to discriminant analysis. Conceptually, this is at the opposite end of the spectrum from the 'traditional plus diagnostics' approach – instead it uses a single fit which is proof against arbitrary outliers whose number is below a set threshold. The price paid for this protection is some loss of statistical efficiency if there are no outliers in the data. If protection is sought against only a few outliers (as will be the case in most applications), then the loss of efficiency will be small. A 'feasible solution' algorithm for the MWCD appears to be able to give the optimum with high probability using only modest computations.

The proposal is illustrated by a reanalysis of two standard data sets from the literature. One data set shows a substantial improvement in classification accuracy from using the robust approach. In the other, while the MWCD does not give any appreciable improvement in performance over the classical approach, it seems to work at least as well.

We see this high breakdown approach as a useful adjunct to discriminant analysis. Where a data set does not show a marked difference in the determinants of the covariance matrix of the full data set and that given by the MWCD criterion, then there is likely little danger in using the traditional approaches. Where the MWCD determinant is substantially smaller than that of the full sample within-group covariance matrix, then the data set is at risk of poor classification rules created by the outliers. It is in this circumstance that the high breakdown method solves a potentially serious problem.

# 6 References


Butler, R. W., Davies, P. L., and Jhun, M., (1993), "Asymptotics for the minimum covariance determinant estimator," *The Annals of Statistics*, **21**, 1385–1400.

Chork, C. Y. and Rousseeuw P. J. (1992), "Integrating a high breakdown option into discriminant analysis in exploration geochemistry," *Journal of Geochemical Exploration*, **43**, 191–203.

Cook, R. D. and Hawkins, D. M. (1990), Discussion of "Unmasking multivariate outliers and leverage points," by P.J. Rousseeuw and B. van Zomeren, *Journal of the American Statistical Association*, **85**, 640–644.

Habbema, J.D.F., Hermans, J., and van den Broek, K. (1974). "A stepwise discriminant analysis program using density estimation," *Compstat 1974, Proceedings Computational Statistics.* Vienna: Physica-Verlag, pp. 101–110.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach based on Influence Functions*, New York: Wiley.

Hawkins, D. M. (1980), *Identification of Outliers*, London: Chapman and Hall.

Hawkins, D. M. (1981), "A new test for multivariate normality and homoscedasticity," *Technometrics*, **23**, 105–110.

Hawkins, D. M. (1994), "The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data," *Computational Statistics and Data Analysis*, **17**, 197–210.

Hawkins, D. M. (1993), "A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data," *Computational Statistics*, **9**, 95–107.

Lopuhaä, H. P. (1991), "Breakdown point and asymptotic properties of multivariate S-estimators and $\tau$-estimators: a summary," in *Directions in Robust Statistics and Diagnostics, Part I*, eds. W. Stahel and S. Weisberg, New York: Springer Verlag, pp. 167–182.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Reaven, G.M. and Miller, R.G. (1979), "An attempt to define the nature of chemical diabetes using a multidimensional analysis," *Diabetologia*, **16**, 17–24.

Rousseeuw, P. J. and Leroy, A. M. (1986), *Robust Regression and Outlier Detection*, New York: Wiley.

Tyler, D. E. (1991), "Some issues in the robust estimation of multivariate location and scatter, in *Directions in Robust Statistics and Diagnostics, Part I*, eds. W. Stahel and S. Weisberg, New York: Springer Verlag, pp. 327–336.

**Table 1. Asymptotic bias correction factors (first line) and location estimator efficiencies (second line) of the MCD estimator of a common covariance matrix**

| | Coverage | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 |
| 1 | 0.176 | 0.215 | 0.259 | 0.310 | 0.369 | 0.438 | 0.521 | 0.623 |
| | 0.097 | 0.129 | 0.168 | 0.217 | 0.276 | 0.350 | 0.442 | 0.561 |
| 2 | 0.347 | 0.389 | 0.435 | 0.484 | 0.538 | 0.598 | 0.665 | 0.744 |
| | 0.191 | 0.233 | 0.283 | 0.339 | 0.403 | 0.478 | 0.565 | 0.670 |
| 3 | 0.446 | 0.486 | 0.529 | 0.573 | 0.621 | 0.674 | 0.731 | 0.797 |
| | 0.245 | 0.292 | 0.344 | 0.401 | 0.466 | 0.539 | 0.622 | 0.717 |
| 4 | 0.511 | 0.549 | 0.588 | 0.629 | 0.673 | 0.719 | 0.770 | 0.828 |
| | 0.281 | 0.329 | 0.382 | 0.441 | 0.505 | 0.576 | 0.655 | 0.745 |
| 5 | 0.558 | 0.594 | 0.630 | 0.668 | 0.708 | 0.751 | 0.797 | 0.849 |
| | 0.307 | 0.356 | 0.410 | 0.468 | 0.531 | 0.601 | 0.677 | 0.764 |
| 6 | 0.594 | 0.627 | 0.662 | 0.697 | 0.734 | 0.774 | 0.816 | 0.864 |
| | 0.326 | 0.376 | 0.430 | 0.488 | 0.551 | 0.619 | 0.694 | 0.777 |
| 7 | 0.622 | 0.654 | 0.686 | 0.720 | 0.755 | 0.792 | 0.831 | 0.875 |
| | 0.342 | 0.392 | 0.446 | 0.504 | 0.566 | 0.633 | 0.707 | 0.788 |
| 8 | 0.645 | 0.676 | 0.707 | 0.738 | 0.771 | 0.806 | 0.843 | 0.884 |
| | 0.355 | 0.405 | 0.459 | 0.517 | 0.578 | 0.645 | 0.717 | 0.796 |
| 9 | 0.665 | 0.694 | 0.723 | 0.754 | 0.785 | 0.818 | 0.853 | 0.892 |
| | 0.365 | 0.416 | 0.470 | 0.528 | 0.589 | 0.654 | 0.725 | 0.803 |
| 10 | 0.681 | 0.709 | 0.738 | 0.766 | 0.796 | 0.828 | 0.861 | 0.898 |
| | 0.375 | 0.425 | 0.479 | 0.536 | 0.597 | 0.662 | 0.732 | 0.808 |
| 11 | 0.695 | 0.722 | 0.750 | 0.778 | 0.806 | 0.836 | 0.868 | 0.903 |
| | 0.382 | 0.433 | 0.487 | 0.544 | 0.605 | 0.669 | 0.738 | 0.813 |
| 12 | 0.708 | 0.734 | 0.760 | 0.787 | 0.815 | 0.844 | 0.874 | 0.908 |
| | 0.389 | 0.440 | 0.494 | 0.551 | 0.611 | 0.675 | 0.743 | 0.817 |
| 13 | 0.719 | 0.744 | 0.770 | 0.796 | 0.823 | 0.850 | 0.880 | 0.912 |
| | 0.396 | 0.447 | 0.500 | 0.557 | 0.617 | 0.680 | 0.748 | 0.821 |
| 14 | 0.729 | 0.754 | 0.778 | 0.803 | 0.829 | 0.856 | 0.885 | 0.915 |
| | 0.401 | 0.452 | 0.506 | 0.562 | 0.622 | 0.685 | 0.752 | 0.824 |
| 15 | 0.738 | 0.762 | 0.786 | 0.810 | 0.835 | 0.861 | 0.889 | 0.919 |
| | 0.406 | 0.457 | 0.511 | 0.567 | 0.626 | 0.689 | 0.755 | 0.827 |
| 20 | 0.773 | 0.794 | 0.815 | 0.836 | 0.858 | 0.881 | 0.905 | 0.931 |
| | 0.425 | 0.476 | 0.530 | 0.585 | 0.644 | 0.705 | 0.769 | 0.838 |
| 25 | 0.796 | 0.816 | 0.835 | 0.854 | 0.874 | 0.894 | 0.916 | 0.939 |
| | 0.438 | 0.489 | 0.543 | 0.598 | 0.655 | 0.715 | 0.778 | 0.845 |

**Table 2. Full-sample summary statistics for the diabetes data set**

| Grp | size | means | | |
|---|---|---|---|---|
| 1 | 76 | 355.9 | 172.6 | 118.6 |
| 2 | 36 | 481.5 | 288.0 | 199.4 |
| 3 | 33 | 1044.0 | 106.0 | 318.9 |
| | | Pooled covariance matrix | | |
| | | 23900 | –3539 | 4261 |
| | | –3539 | 10610 | 1745 |
| | | 4261 | 1745 | 4821 |

Log generalized variance=27.4611

**Table 3. Log generalized variance for various coverages**

| Coverage(%) | Log generalized variance |
|---|---|
| 100 | 27.46 |
| 95 | 26.46 |
| 90 | 25.72 |
| 85 | 25.03 |
| 80 | 24.13 |
| 75 | 23.62 |
| 70 | 23.19 |
| 65 | 23.03 |
| 60 | 22.68 |

**Table 4. Comparative cross-validated allocations**

| Coverage (%) | Both correct | Robust correct, nonrobust incorrect | Nonrobust correct, robust incorrect | Both incorrect |
|---|---|---|---|---|
| 95 | 118 | 0 | 0 | 27 |
| 90 | 118 | 7 | 0 | 20 |
| 85 | 118 | 10 | 0 | 17 |
| 80 | 115 | 12 | 3 | 15 |
| 75 | 115 | 12 | 3 | 15 |
| 70 | 114 | 7 | 4 | 20 |
| 65 | 115 | 8 | 3 | 19 |
| 60 | 116 | 10 | 2 | 17 |

**Table 5. Summary statistics of the covered cases at 85% coverage**

| Grp | size | | means | |
|---|---|---|---|---|
| 1 | 76 | 355.9 | 172.6 | 118.6 |
| 2 | 31 | 481.8 | 250.2 | 196.6 |
| 3 | 18 | 888.3 | 113.5 | 290.1 |
| | | Bias-corrected covariance matrix | | |
| | | 7771 | –182 | 2318 |
| | | –182 | 8377 | 3668 |
| | | 2318 | 3668 | 8095 |

**Table 6. Details of feasible solutions**

| Coverage(%) | No. of solutions | Percentage converging to best |
|---|---|---|
| 95 | 1 | 100 |
| 90 | 2 | 98 |
| 85 | 1 | 100 |
| 80 | 2 | 99 |
| 75 | 2 | 97 |
| 70 | 8 | 62 |
| 65 | 15 | 65 |
| 60 | 8 | 91 |

**Table 7. Full-sample summary statistics for the hemophilia data set**

| Grp | size | | means | |
|---|---|---|---|---|
| 1 | 30 | | –0.1347 | –0.0078 |
| 2 | 45 | | –0.3079 | –0.0060 |
| | | Pooled covariance matrix | | |
| | | | 0.0226 | 0.0154 |
| | | | 0.0154 | 0.0216 |

Log generalized variance=–8.2905

**Table 8. Summary statistics of the covered cases at 75% coverage**

| Grp | size | means | |
|---|---|---|---|
| 1 | 23 | –0.1290 | –0.0603 |
| 2 | 33 | –0.3291 | –0.0024 |
| Bias-corrected covariance matrix | | | |
| | | 0.0158 | 0.0016 |
| | | 0.0016 | 0.0316 |

Log generalized variance=–9.5002

**Table 9 Comparative cross-validated allocations**

| Both correct | Robust correct, nonrobust incorrect | Nonrobust correct, robust incorrect | Both incorrect |
|---|---|---|---|
| 59 | 4 | 3 | 9 |