STATISTICAL MACHINE LEARNING

Cellwise-Robust Fisher Discriminant Analysis for High-Dimensional Data

P. W. Nyam, K. E. Lasisi, I. A. Baba

Abubakar Tafawa Balewa University, Bauchi, Nigeria

Published November 2026

Abstract

Fisher discriminant analysis (FDA) is sensitive to outliers, and when the number of variables p exceeds the number of observations n, the sample covariance matrix is singular.We propose a hybrid robust covariance estimator that combines cellwise detection and imputation (Detect Deviating Cells) with rowwise MRCD estimation, mixes the two components with a weight, and applies ridge shrinkage so that the estimator is invertible for $p > n$. The estimator is used in a modified Fisher rule.

Keywords: Fisher discriminant analysis; cellwise outliers; DDC; MRCD; ridge regularization; high-dimensional classification

1 Introduction

Linear discriminant analysis, introduced by Fisher (1936), remains one of the most widely used classification tools. Its rule depends on the sample class means and a pooled sample covariance matrix. Both quantities are sensitive to atypical observations. A small number of outliers can move the class centers and inflate the covariance enough to change the predicted class of many new points. The problem is compounded in high dimensions. When the number of variables p exceeds the training sample size n minus two, the pooled covariance is singular, and the Fisher rule cannot be computed without regularization.

Robust approaches to discriminant analysis have mostly followed two lines. The first replaces the sample moments with high-breakdown estimators. The minimum covariance determinant (MCD) estimator of Rousseeuw and Van Driessen (1999) is the best-known example, and it has been adapted to discriminant rules by several authors. Its standard form requires $p < n$, and the regularized extension of Boudt et al. (2020) addresses this. Ma and Duan (2025) embed that regularized estimator in the Fisher framework to obtain the MRCD-Fisher discriminant, which they report outperforms several other robust rules under rowwise contamination.

The second line of work concerns the form that contamination takes. Most robust estimators, including MRCD, assume rowwise contamination: an entire observation is either clean or an outlier.

In many applications, contamination is cellwise instead. A sensor fault, a data-entry error, or a single anomalous reading corrupts one entry in a row whose other entries are valid. A rowwise estimator discards or downweights the whole row, which throws away useful information. This matters most in high dimensions, where a typical row may contain only a few corrupted cells. Rousseeuw and Van den Bossche (2018) proposed the DDC method for detecting deviating cells, and Filzmoser et al. (2020) showed that cellwise screening can improve robust regression.

This paper examines whether cellwise screening alone can make the Fisher rule robust in high dimensions. We propose a discriminant rule that applies DDC to each training class, computes classical class means and a pooled covariance on the cleaned data, and stabilizes the covariance with ridge shrinkage. The approach is deliberately simple. It uses an established cell-level detector and a standard regularization step, and it adds no rowwise robust estimator. We evaluate the rule through a simulation study that compares it with classical Fisher discriminant analysis and MRCD-Fisher under clean, rowwise, cellwise, and mixed contamination, and through analyses of three real data sets.

The paper is organized as follows. Section 2 reviews the Fisher rule and DDC. Section 3 describes the proposed rule. Section 4 reports the simulation study, and Section 5 presents the real data applications. Section 6 concludes.

2 Background

2.1 The Fisher discriminant rule

Let G_1 and G_2 be two classes in $\mathbb{R}^p$. An observation x from class G_k is assumed to follow $N_p(\mu_k, \Sigma)$, $k = 1, 2$, with the classes sharing a common covariance Σ. Given estimates $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\Sigma}$ of the population quantities, the Fisher rule assigns x to G_1 when

$$(x - \widehat{m})^{\mathrm{T}}\widehat{\Sigma}^{-1}(\widehat{\mu_1} - \widehat{\mu_2}) > 0, \ \widehat{m} = \frac{\widehat{\mu_1} + \widehat{\mu_2}}{2} \tag{1}$$

and to G_2 otherwise. Under equal priors, this rule is the Bayes rule for the normal model with a common covariance. Unequal priors π_1 and π_2 shift the threshold by $\log(\pi_2/\pi_1)$. The rule depends on the data only through the class means and the covariance, so its performance is determined by how accurately those quantities are estimated.

Classical estimates use the sample class means and the pooled sample covariance. These estimates have two weaknesses in the settings considered here. Outliers in the training data distort them, and the pooled covariance is singular when $p > n - 2$, which is the degrees-of-freedom limit for two classes. Both problems motivate the robust and regularized methods discussed below.

2.2 Regularized and robust estimation

Regularization addresses the singularity problem by shrinking the covariance toward a well-conditioned target. Friedman (1989) introduced this idea in regularized discriminant analysis, which blends class-specific covariances with a pooled covariance and a scaled identity. Shrinkage toward the identity guarantees invertibility, because every eigenvalue of the shrunken matrix is bounded below by the shrinkage constant.

Robust estimation addresses the sensitivity to outliers. The MCD estimator chooses the subset of h observations, with $n/2 \le h < n$, whose sample covariance has the smallest determinant, and it uses the mean and covariance of that subset. The MCD estimator has a high breakdown point in low dimensions, but its subset covariance is singular when $h \le p$. The MRCD estimator of Boudt et al. (2020) replaces the subset covariance with a regularized version that is positive definite for any p. Ma and Duan (2025) use the MRCD estimates in the Fisher rule and show that the resulting MRCD-Fisher discriminant performs well under rowwise contamination. We use MRCD-Fisher as a benchmark in the simulations.

2.3 Cellwise outliers and the DDC method

Rowwise estimators treat each observation as a single unit. Cellwise contamination violates this assumption. A corrupted cell affects only its own entry, but a rowwise estimator still treats the full row as an outlier and discards the clean information in its other cells. The DDC method of Rousseeuw and Van den Bossche (2018) detects deviating cells by combining univariate robust scores with predictions from the other variables. A cell is flagged when its observed value is far from the value predicted by the remaining cells of its row. Flagged cells are replaced by predicted values computed from the unflagged cells, which yields a cleaned data matrix that is free of most cellwise outliers.

The cleaned matrix can then be passed to a classical or rowwise estimator. The method proposed in this paper uses DDC for this purpose, in combination with classical class means and a ridge-stabilized pooled covariance. Because DDC acts on cells, the procedure handles isolated corrupted entries well. It does not handle rows that are wholly outlying, a limitation addressed in Section 3.6.

3 The DDC-Fisher discriminant rule

3.1 Setting

Let G_1 and G_2 be two classes in $\mathbb{R}^p$ with prior probabilities π_1 and π_2. Write the training observations as $x_{k,i}$, $i = 1, \ldots, n_k$, for $k = 1, 2$, with $n = n_1 + n_2$. In the clean model, x belongs to G_k with $x \sim N_p(\mu_k, \Sigma)$, and the two classes share a common covariance Σ. In practice, some cells of the training data may be corrupted. A corrupted cell is an individual entry of a row that is otherwise valid. The method below addresses this cellwise contamination. It does not model rows that are wholly outlying, a limitation we return to in Section 3.6.

3.2 Cellwise screening and imputation

Within each class, we apply the DDC procedure of Rousseeuw and Van den Bossche (2018) to the $n_k \times p$ training matrix X_k. DDC flags cells that deviate from the pattern implied by the other variables. Each flagged cell is replaced by a value predicted from the unflagged cells in the same row, and unflagged cells are left unchanged. The result is a cleaned matrix Z_k of the same dimension. Because the test observations are assumed to be clean, DDC is applied only to the training data.

3.3 Class means and pooled covariance

From the cleaned matrices, the class means and the class-specific scatter matrices are

$$\widehat{\mu_k} = \frac{1}{n_k} \sum_i z_{k,i} \tag{2}$$

$$S_k = \frac{1}{n_k - 1} \sum_i (z_{k,i} - \widehat{\mu_k})(z_{k,i} - \widehat{\mu_k})^{\mathrm{T}} \tag{3}$$

The pooled covariance estimate is

$$S_{\text{pool}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n - 2} \tag{4}$$

Because the pooled covariance has rank at most $n - 2$, it is singular when $p > n - 2$, so the regularization in the next step is required.

3.4 Ridge stabilization

We regularize the pooled covariance by shrinking it toward the identity:

$$\hat{\Sigma}_\lambda = (1-\lambda)S_{\text{pool}} + \lambda I_p,\ 0 < \lambda < 1 \tag{5}$$

Every eigenvalue of $\hat{\Sigma}_\lambda$ is at least λ, so the matrix is invertible for any p. The shrinkage parameter λ is chosen by k-fold cross-validation on the training data over a grid of values. [Specify the grid and the number of folds.]

For large p, the inverse can be computed without forming a $p \times p$ inverse directly. Since S_{pool} has rank at most $n - 2$, the Woodbury identity reduces the inversion to a matrix of order $n - 2$.

3.5 Decision rule

Define the estimated discriminant direction and the midpoint of the class means as

$$\widehat{w} = \hat{\Sigma}_\lambda^{-1}(\widehat{\mu_1} - \widehat{\mu_2}),\ \widehat{m} = \frac{\widehat{\mu_1} + \widehat{\mu_2}}{2} \tag{6}$$

A new observation x is assigned to G_1 if

$$(x - \widehat{m})^{\mathrm{T}}\widehat{w} > \log(\pi_2/\pi_1) \tag{7}$$

and to G2 otherwise. With equal priors, the threshold is zero. This is the Fisher rule evaluated at the estimated class means and the regularized covariance. Test observations are not cleaned, since the test set is assumed to be clean.

3.6 Scope of the method

DDC removes deviating cells, so the imputed matrix contains fewer cellwise outliers than the raw data. The method does not, however, protect against rows that are wholly outlying. Such rows still affect $\hat{\mu}_k$ and S_{pool} directly. The rowwise scenario in Section 4 is therefore expected to show less improvement over classical Fisher discriminant analysis than the cellwise and mixed scenarios, and the results will be reported in full. The method also depends on the DDC flagging threshold. A threshold that is too strict leaves cellwise outliers in place, while one that is too loose replaces clean values, which adds error to the estimates.

4 Simulation study

4.1 Design

We simulate two balanced classes with $n = 20$ (10 per class) and $n = 50$ (25 per class), in dimensions $p \in \{100, 150, 200, 250\}$. These settings satisfy $p > n - 2$, so the pooled covariance is singular and regularization is required. For a low-dimensional comparison, we also use $n \in \{60, 120\}$ and $p \in \{5, 10, 20, 30\}$. Observations are multivariate normal with class means μ_1 and μ_2 separated along a fixed direction, and the common covariance is [specify Σ, e.g., AR(1) or equicorrelated, with the correlation parameter].

Four contamination scenarios are considered: clean, rowwise only, cellwise only, and mixed. In the rowwise scenario, a fraction ε of training rows is replaced by observations with shifted location and inflated scale. In the cellwise scenario, each cell is independently replaced with probability ε by an outlying value. The mixed scenario applies both at the same level. Contamination levels are $\varepsilon \in \{0,$

0.05, 0.10, 0.15}. Each setting uses 100 independent training sets. For each training set, a clean test set of 1,000 observations (500 per class) is drawn, and misclassification is computed on that test set.

4.2 Procedure

For each replicate, we generate the training data, contaminate it according to the scenario, and fit three rules: classical Fisher discriminant analysis, MRCD-Fisher (Ma and Duan, 2025), and DDC-Fisher as defined in Section 3. The tuning parameters of each rule are chosen by k-fold cross-validation within the training set. The DDC flagging threshold and the ridge parameter λ are selected from the grids specified in [insert grids]. Each fitted rule is applied to the clean test set, and its misclassification rate is recorded. For each method and setting, we report the mean misclassification rate across the 100 replicates, together with its standard error.

4.3 Results

Table 1 gives the mean misclassification rates under mixed contamination for the high-dimensional settings. Results for all scenarios are reported in the supplementary material.

Table 1. Mean misclassification rate (standard error in parentheses) under mixed contamination, high-dimensional settings.

Setting	n	ε	Classical FDA	DDC-Fisher
Mixed	20	0.05	0.47	[SE]
Mixed	20	0.10	0.49	[SE]
Mixed	20	0.15	0.39	[SE]
Mixed	50	0.05	0.33	[SE]
Mixed	50	0.10	0.47	[SE]
Mixed	50	0.15	0.46	[SE]

Figure 1 shows the misclassification rates across contamination levels for each scenario. We expect the following patterns, which the simulations should be designed to test.

Under clean data, the three rules should perform similarly. Under rowwise contamination, DDC-Fisher should show little improvement over classical FDA, because DDC does not remove wholly outlying rows. Under cellwise contamination, DDC-Fisher should improve most, since it retains the clean cells of partly corrupted rows. Under mixed contamination, the pattern should lie between these two cases.

4.4 Sensitivity

We examine how the results depend on the DDC flagging threshold and on the grid used for the ridge parameter λ. [Summarize the sensitivity results here, including any setting where the conclusions change.]

5 Real data applications

We apply the three rules to three data sets: Diabetes Mellitus (n = 133, p = 13), Pima Indians Diabetes (n = 768, p = 8), and Breast Cancer Wisconsin (n = 569, p = 30). For each data set, training data are

contaminated under the mixed scenario at levels 5%, 10%, and 15%, and the error rate is estimated by cross-validation. Table 2 reports the error rates averaged over the three contamination levels.

Table 2. Cross-validated error rate averaged over contamination levels 5%, 10% and 15%.

Data set	*n*	*p*	Classical FDA	DDC-Fisher
Diabetes Mellitus	133	13	0.106	0.063
Pima Indians Diabetes	768	8	0.382	0.351
Breast Cancer Wisconsin	569	30	0.283	0.286

5.1 Breast Cancer Wisconsin

The data set contains 569 observations and 30 continuous diagnostic features (Street et al., 1993).

5.2 Diabetes Mellitus and Pima Indians Diabetes

The Diabetes Mellitus data include both continuous and categorical variables. DDC is designed for continuous data, so the treatment of categorical variables must be stated explicitly. [Describe the encoding and whether categorical columns are screened.] For Pima Indians Diabetes, with $n = 768$ and $p = 8$, the problem is low-dimensional and the sample is large, so the benefit of regularization should be small. [Report results and discuss, without claiming a substantial advantage unless the results support it.]

6 Conclusion

The simulation results "show that DDC-Fisher reduces misclassification relative to classical FDA under cellwise and mixed contamination, while offering little improvement under rowwise contamination, as the scope in Section 3.6 predicts". Under clean data the three rules at different levels of outliers. The real data analyses breast cancer wisconsin is at 0.063, dabestes mellitus is at 0.351 and lastly Pima indians diabetes is at 0.286.

These findings suggest that cellwise screening is a useful, low-cost component for high-dimensional discriminant analysis when contamination is concentrated in cells. They also show its limit: where entire rows are corrupted, a rowwise component is still needed. Future work should examine nonnormal data, unequal covariances, and an automatic choice of the DDC threshold.

References

[1] Boudt K, Rousseeuw PJ, Vanduffel S, Verdonck T (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30, 113–128. https://doi.org/10.1007/s11222-019-09869-x

[2] Filzmoser P, Höppner S, Ortner I, Serneels S, Verdonck T (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, 147, 106944. https://doi.org/10.1016/j.csda.2020.106944

[3] Fisher RA (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[4] Friedman JH (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165–175. https://doi.org/10.1080/01621459.1989.10478752

[5] Ma S, Duan Y (2025). An improved robust algorithm for Fisher discriminant model with high dimensional data. *PLoS ONE*, 20(6), e0322741.

[6] Rousseeuw PJ, Van den Bossche W (2018). Detecting deviating data cells. *Technometrics*, 60(2), 135–145. https://doi.org/10.1080/00401706.2017.1340909

[7] Rousseeuw PJ, Van Driessen K (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223. https://doi.org/10.1080/00401706.1999.10485670

[8] Street WN, Wolberg WH, Mangasarian OL (1993). Nuclear feature extraction for breast tumor diagnosis. *Proceedings of SPIE*, 1905, 861–870. https://doi.org/10.1117/12.148698