Robust Estimation for Fisher Discriminant Analysis

Steven Horstink

Bachelor thesis Applied Mathematics

University of Twente

June 29, 2018

Supervisors: Dr. Ir. J. Goseling (SOR)
Dr. Ir. L.J. Spreeuwers (DMB)

Abstract

Fisher Linear Discriminant Analysis (LDA) is a well-known classification method, but it is also well-known for not being robust against outliers. This paper investigates the uses of two methods for data classification including outliers. One method alleviates data sensitivity by incorporating data uncertainty and subsequently optimizes the worst-case scenario of the Fisher discriminant ratio, which appears to be ineffective. The use of the second method does seem to be effective. It directly attempts to remove outliers by removing those points that lie furthest from the sample mean in the Mahalanobis distance sense. Additionally, this paper provides a proof for a general tolerance ellipsoid for multivariate normally distributed data which is used in the second method. This technique is also well-known and a rather obvious one, yet most papers do not provide a general proof for this concept.

1 Introduction

Nowadays there exist many statistical classification algorithms that attempt to identify to which class a new observation $\boldsymbol{s}$ belongs, given $\{C_1, \ldots, C_K\}$, a set of K classes. This can be done in a wide variety of ways that can be condensed into three different approaches [1]: in decreasing order of complexity, one could determine $p(C_k, \boldsymbol{s})$ and find $p(C_k \mid \boldsymbol{s})$ using Bayes' rule, called *generative modelling*, one could directly compute $p(C_k \mid \boldsymbol{s})$, called *discriminative modelling*, or one could simply find a *discriminative function* $f(\boldsymbol{s})$ that directly maps $\boldsymbol{s}$ onto a class label. Regarding the first two, classification can be performed after obtaining $p(C_k \mid \boldsymbol{s})$ for every class by using the *maximum likelihood discriminant rule*, which assigns $\boldsymbol{s}$ to C_j if $p(C_j \mid \boldsymbol{s}) \geq p(C_k \mid \boldsymbol{s})$ for all k [2].

Both generative and disciminative models are instances of supervised learning. In supervised learning, the discriminant rule is based on available data, called the training set. If this available data is corrupted in the sense that it contains outliers, the perceptions of $p(C_k \mid \boldsymbol{s})$ are easily influenced, possibly producing poor classification results. Outliers in this paper are considered to be data points that are classified as belonging to a certain class but do not have the same distribution as that class. For several classification algorithms inherent robust methods have been constructed, *e.g.* [3–5] for Principal Components Analysis, which is in essence a dimensionality reducer but can be used as classifier, or [6, 7], which are direct applications of Fisher LDA to face recognition. Robustness can also imply robustness against a small sample size, for which [8, 9] provide a solution. Extrinsic robust methods can also be used, *e.g.* [10], which describes general robustness of estimates. There exist no methods for Fisher LDA specifically that are robust against outliers.

Fisher LDA is a generative classifier, also the most well-known *linear* classifier. The goal of an LDA is preprocessing the data by projecting a data set of M-dimensional samples onto a smaller subspace while maintaining the class-discriminatory information. The popularity of LDAs lies in their simplicity and computationally inexpensiveness. Originally, Ronald A. Fisher introduced the concept of transforming two classes of M-dimensional data to 1-dimensional data using a discriminant $\boldsymbol{w}$ that maximizes class separation and minimizes within class covariance, hence the name [11]. By now, it has been extended to $K > 2$ classes and non-linear classification [12–14], although this paper only attempts to classify a new observation to $K = 2$ multivariate normally distributed classes. Considering $K > 2$ classes would redefine the definition of the Fisher discriminant ratio (12) and its derivation (3).

This paper will investigate the use of two methods of estimating the means and covariance for Fisher LDA, which are called *the worst-case estimates* and *the* $\boldsymbol{t}_{M,p}$ *estimates*. The first of the two methods is inherent to Fisher LDA and is introduced in [15], which claims it alleviates data sensitivity by incorporating data uncertainty and subsequently optimizes the worst-case scenario of the Fisher discriminant ratio (12). This paper demonstrates that this method is ineffective for using it as a robust method against outliers. Following the poor performance of this method against outliers come the alternative $\boldsymbol{t}_{M,p}$ estimates. The tolerance ellipsoid, defined by a number of dimensions M and tolerance parameter p, encompasses a fraction p of n points in M-dimensional space as n goes to infinity. A common type of outliers, i.e. outliers that lie further from the mean in relation to the variance, can be spotted and removed. This does appear to be an effective method. Many articles discuss the use of this tolerance ellipsoid but do not provide a clear definition and proof, *e.g.* [16–19]. Therefore, this paper also provides a proof for the construction of the $\boldsymbol{t}_{M,p}$ estimates for multivariate normally distributed data.

2 Problem statement

Suppose that there are two classes X and Y in an M-dimensional space $\mathbb{R}^{M\times 1}$, assumed to be multivariate normally distributed. Of both classes we obtain samples/observations as column vectors, denoted as $\boldsymbol{x}$ and $\boldsymbol{y}$, called our sample set or training set. Each dimension in these M-dimensional vectors contains specific information about the sample, of which the total M-dimensional information will eventually dictate to which class the sample belongs. Therefore, given a new sample $\boldsymbol{s}$ drawn from either of the two distributions of our classes X and Y, it is the task of a classifier to tell us to which of the two classes the new sample $\boldsymbol{s}$ belongs.

First, classification will be discussed in Section 2.1 after which Fisher LDA will be introduced in Section 2.2. The main problem that this paper discusses is as follows. The discriminant $\boldsymbol{w}$ is computed in (3) using the covariance matrices $\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y$ and means $\boldsymbol{\mu}_x, \boldsymbol{\mu}_x$ of both classes. Since we only have a sample set to represent our classes, we must estimate the covariances and means based on the sample set. The regular non-robust estimates are the sample covariance and sample mean, which for class X would be

$$\hat{\boldsymbol{\mu}}_x = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i, \qquad\qquad \hat{\boldsymbol{\Sigma}}_x = \frac{1}{N-1}(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_x)(\boldsymbol{X} - \hat{\boldsymbol{\mu}}_x)^T,$$

where $\boldsymbol{X}$ is a matrix with samples $\boldsymbol{x}_i$ as its columns, $i = 1, \ldots, N$. However, should the sample set contain outliers, then the sample mean and sample covariance are easily influenced, resulting in possibly poor classification. An example will be given in Section 2.3.

Therefore, this paper investigates in Section 3 the use of the two methods mentioned before and answer the question whether or not using the methods improves success rates. Success rates are found by drawing a test set of 1.000 samples from the distributions of both classes. We let the classifier do

its work and base the success rate on the fraction of test samples correctly classified by the classifier. We obtain two success rates: one for class X and one for class Y. We compute the average of these two success rates and let that be our final success rate. The results of the influence of the methods on the success rates are given in Section 4.

2.1 Classification

We need a method that assigns a newly drawn sample $\boldsymbol{s}$ to the class it belongs to. The maximum likelihood discriminant rule, which assigns $\boldsymbol{s}$ to C_j if $p\left(C_j \mid \boldsymbol{s}\right) \geq p\left(C_k \mid \boldsymbol{s}\right)$ for all k, is an admissible discriminant rule [2]. This means that there is no better discriminant rule. In the case of two classes, we assign $\boldsymbol{s}$ to X if

$$\frac{P\left[X \mid \boldsymbol{s}\right]}{P\left[Y \mid \boldsymbol{s}\right]} > 1.$$

Let us first take a look at $P\left[X \mid \boldsymbol{s}\right]$, which is the probability of class X being referenced to by the sample $\boldsymbol{s}$. Using Bayes' rule, we derive

$$P\left[X \mid \boldsymbol{s}\right] = \frac{p\left(\boldsymbol{s} \mid X\right) P\left[X\right]}{p\left(\boldsymbol{s}\right)},$$

where $p\left(\boldsymbol{s} \mid X\right)$ is the probability density of $\boldsymbol{s}$ originating from the distribution of X, also denoted as $p_X\left(\boldsymbol{s}\right)$. The probability $P\left[X\right]$ is the probability that any random sample originates from X, which depends on your prior knowledge of your two classes. In this paper we assume $P\left[X\right] = P\left[Y\right]$, but these values could be approximated as the number of observations of one class divided by the total number of observations. The probability density of $\boldsymbol{s}$ originating from either of the two distributions of the classes X and Y is given by $p\left(\boldsymbol{s}\right) = p_X\left(\boldsymbol{s}\right) P\left[X\right] + p_Y\left(\boldsymbol{s}\right) P\left[Y\right]$. We do the same for class Y. Now, the maximum likelihood discriminant rule dictates that we assign $\boldsymbol{s}$ to X if

$$\frac{P\left[X \mid \boldsymbol{s}\right]}{P\left[Y \mid \boldsymbol{s}\right]} = \frac{p_X\left(\boldsymbol{s}\right) P\left[X\right]}{p\left(\boldsymbol{s}\right)} \Bigg/ \frac{p_Y\left(\boldsymbol{s}\right) P\left[X\right]}{p\left(\boldsymbol{s}\right)} = \frac{p_X\left(\boldsymbol{s}\right)}{p_Y\left(\boldsymbol{s}\right)} > 1. \tag{1}$$

However, calculating $p_X\left(\boldsymbol{s}\right)$ and $p_Y\left(\boldsymbol{s}\right)$ requires a lot of computational power if the number of dimensions M is large. Therefore, we want to reduce the number of dimensions while preserving the class-discriminatory information.

2.2 Fisher's linear discriminant

Let us define a linear mapping $f : \mathbb{R}^{M\times 1} \to \mathbb{R}$ that takes a sample $\boldsymbol{s} \in \mathbb{R}^{M\times 1}$ as input and outputs the projection of $\boldsymbol{s}$ on a 1-dimensional space,

$$f(\boldsymbol{s}) = \boldsymbol{w}^T \boldsymbol{s}.$$

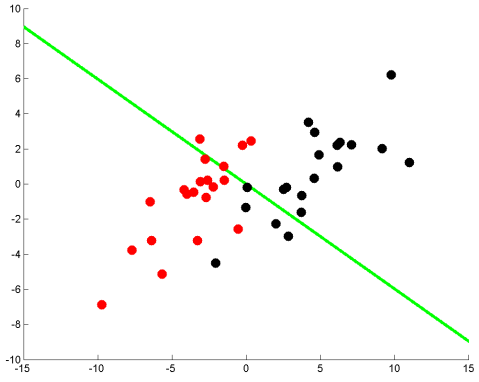
Notice that this is a linear transformation of multivariate normally distributed data, which is again a normal distribution (see appendix, Theorem 1). According to Theorem 1, the mapping of the distribution of class X onto $\mathbb{R}$ yields a univariate normal distribution, such that

$$X_W \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{\mu}, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}) \quad \Longrightarrow \quad p_{X_W}\left(s\right) = \frac{1}{\sqrt{2\pi\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} \exp\left(-\frac{1}{2}\frac{(s-\boldsymbol{w}^T\boldsymbol{\mu})^2}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}\right). \tag{2}$$

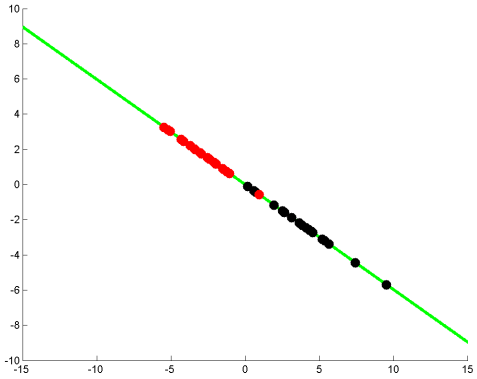
Specifically, we want this linear transformation to optimally seperate our two classes X and Y according to Fisher's linear discriminant (see appendix, Theorem 2). This discriminant is given by

$$\boldsymbol{w} = (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1}(\boldsymbol{\mu_x} - \boldsymbol{\mu_y}), \tag{3}$$

where $\boldsymbol{w} \in \mathbb{R}^{M\times 1}$. By maximizing the Fisher discriminant ratio (12) along the variable $\boldsymbol{w}$, we simultaneously maximize $\boldsymbol{w}^T(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y))^2$ and minimize $\boldsymbol{w}^T(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)\boldsymbol{w}$. Therefore, using the optimal discriminant $\boldsymbol{w}$ yields maximum seperation between the means $\boldsymbol{w}^T\boldsymbol{\mu}_x$ and $\boldsymbol{w}^T\boldsymbol{\mu}_y$ and minimal values for the covariances $\boldsymbol{w}^T\boldsymbol{\Sigma}_x\boldsymbol{w}$ and $\boldsymbol{w}^T\boldsymbol{\Sigma}_y\boldsymbol{w}$.



(a) The line is the visualization of $\boldsymbol{w}$ as the extention of the vector (3).



(b) Samples projected onto $\boldsymbol{w}$.

Figure 1: Projection of 2-dimensional space onto a 1-dimensional space.

Now, we replace $p_X(\boldsymbol{s})$ in (1) by projecting $p_X(\boldsymbol{s})$ on $\mathbb{R}$ using the discriminant (3) and the linear transformation given by (2). Doing so yields a univariate distribution $p_{X_W}(\boldsymbol{s})$ where the class-discriminatory information has been preserved and thus provides us with an accurate representation of $p_X(\boldsymbol{s})$. Then, we find that

$$\frac{P[X \mid \boldsymbol{s}]}{P[Y \mid \boldsymbol{s}]} = \frac{p_{X_W}(\boldsymbol{w}^T\boldsymbol{s})}{p_{Y_W}(\boldsymbol{w}^T\boldsymbol{s})}, \tag{4}$$

which we call our classifier.

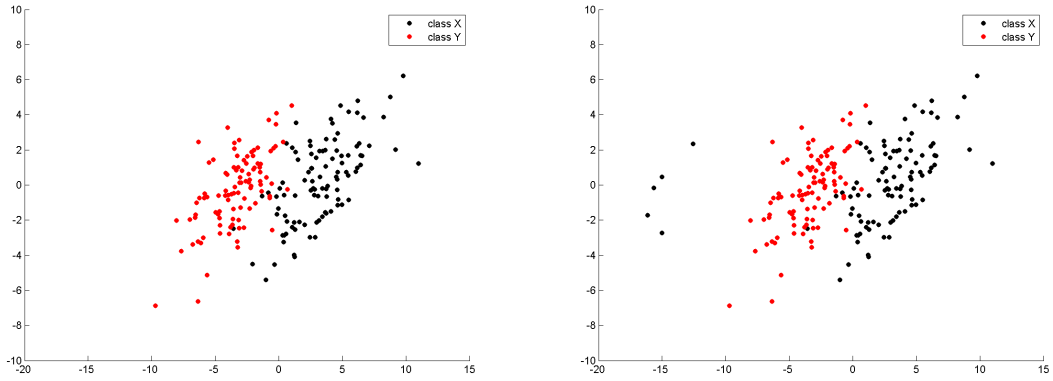
2.3 Numerical example and outliers

To demonstrate the classifier based on Fisher LDA and the consequence of outliers on this classifier, we draw 100 samples from two classes X and Y and call this our sample set. The true means and covariances are

$$\boldsymbol{\mu}_x = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\mu}_y = \begin{bmatrix} -3 \\ 0 \end{bmatrix},$$
$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_x.$$

Classification will be executed once on the sample set and once on the sample set in which 5% of the samples of class X have been replaced with outliers. These outliers will be taken from a multivariate normal distribution with mean and covariance

$$\boldsymbol{\mu} = \begin{bmatrix} -15 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_x.$$



(a) Scatter plot of the sample set.

(b) Scatter plot of the sample set with outliers.

Figure 2: Example of two 2-dimensional classes.

It should be mentioned that some distributions for outliers do not influence the classification success rate much. These are not interesting to consider. Therefore, this distribution for outliers has been chosen somewhat specifically to demonstrate what the influence *could* be.

Classification based on the information given by the sample set and corrupt sample set will now be executed simultaneously. The results of the sample set without outliers will be displayed on the left side and the results of the sample set with outliers on the right side. We begin with a visualization of the two sample sets in Figure 2. An additional visualization of the influence of outliers on the sample covariance is given in Figure 5.

By calculating the sample mean and sample covariance, we have

$$\hat{\boldsymbol{\mu}}_x = \begin{bmatrix} 3.28 \\ 0.28 \end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_y = \begin{bmatrix} -3.38 \\ -0.29 \end{bmatrix}, \qquad \hat{\boldsymbol{\mu}}_x = \begin{bmatrix} 2.35 \\ 0.24 \end{bmatrix}, \quad \hat{\boldsymbol{\mu}}_y = \begin{bmatrix} -3.38 \\ -0.29 \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_x = \begin{bmatrix} 6.76 & 3.88 \\ 3.88 & 5.49 \end{bmatrix}, \qquad \hat{\boldsymbol{\Sigma}}_x = \begin{bmatrix} 22.14 & 4.14 \\ 4.14 & 5.24 \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_y = \begin{bmatrix} 4.31 & 2.66 \\ 2.66 & 4.44 \end{bmatrix}. \qquad \hat{\boldsymbol{\Sigma}}_y = \begin{bmatrix} 4.31 & 2.66 \\ 2.66 & 4.44 \end{bmatrix}.$$

From these estimates we find the discriminant $\boldsymbol{w}$,

$$\boldsymbol{w} = \begin{bmatrix} 0.93 \\ -0.56 \end{bmatrix}. \qquad \boldsymbol{w} = \begin{bmatrix} 0.25 \\ -0.12 \end{bmatrix}.$$

We can now construct our classifier (4) based on our estimates $\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\mu}}_y, \hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$. The success rate is

0.9536 0.8049

3 Analysis

In this section the two methods for robust Fisher LDA will be analysed. First, we will discuss the worst-case method introduced in [15] and see that it cannot be robust against outliers. Next, the $\boldsymbol{t}_{M,p}$ estimates will be introduced by defining its objective and a proof for its construction. Numerical results of using these two methods in the classification process are presented in Section 4.

3.1 Optimizing the Fisher discriminant ratio over the worst-case scenario

Intuitively, [15] attempts to alleviate the sensitivity problem by assuming the, as of yet undefined, worst-case estimation of the means and covariances of X and Y for optimizing the Fisher discriminant ratio (12). This way, Fisher's discriminant is optimized for bad estimations of the means and covariance. The question then arises as to what sort of sensitivity it attempts to counter. This will be discussed later.

Formally, the worst-case scenario is defined to the set of means and covariances $\check{\boldsymbol{\mu}}_x, \check{\boldsymbol{\mu}}_y, \check{\boldsymbol{\Sigma}}_x$ and $\check{\boldsymbol{\Sigma}}_x$ for which (12) is minimal with fixed $\boldsymbol{w}$ and variables $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_x$. After minimizing, we maximize (12) with variable $\boldsymbol{w}$, resulting again in the optimal discriminant (3). This optimization problem is defined as

$$\begin{aligned} \text{minimize} \quad & (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)^{-1} (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y) \\ \text{subject to} \quad & (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y) \in \mathcal{U}. \end{aligned} \tag{5}$$

Here, $\mathcal{U}$ is defined as a convex set established by the constraints

$$\begin{aligned} (\boldsymbol{\mu}_x - \bar{\boldsymbol{\mu}}_x)^T \boldsymbol{P}_x (\boldsymbol{\mu}_x - \bar{\boldsymbol{\mu}}_x) \leq 1, \qquad & ||\boldsymbol{\Sigma}_x - \bar{\boldsymbol{\Sigma}}_x||_F \leq \rho_x, \\ (\boldsymbol{\mu}_y - \bar{\boldsymbol{\mu}}_y)^T \boldsymbol{P}_y (\boldsymbol{\mu}_x - \bar{\boldsymbol{\mu}}_y) \leq 1, \qquad & ||\boldsymbol{\Sigma}_y - \bar{\boldsymbol{\Sigma}}_y||_F \leq \rho_y, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{P}_x &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}^{-1} / M, & \rho_x &= \max_{j=1,\ldots} \left(||\boldsymbol{\Sigma}_{x_j} - \bar{\boldsymbol{\Sigma}}_x||_F \right), \\ \boldsymbol{P}_y &= \boldsymbol{\Sigma}_{\boldsymbol{\mu}_y}^{-1} / M, & \rho_y &= \max_{j=1,\ldots} \left(||\boldsymbol{\Sigma}_{y_j} - \bar{\boldsymbol{\Sigma}}_y||_F \right). \end{aligned}$$

Through bootstrapping [20] we obtain 100 new sets of the data set and from those resamples we obtain a set of 100 sample means and sample covariances for X and Y. From these sets we compute the nominal means and covariances, $\bar{\boldsymbol{\mu}}_x, \bar{\boldsymbol{\mu}}_y, \bar{\boldsymbol{\Sigma}}_x$ and $\bar{\boldsymbol{\Sigma}}_x$, as pointwise averages. From the set of means we also compute its covariances $\boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\mu}_y}$. [15] claims that the constraint $(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \boldsymbol{P} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \leq 1$ corresponds to a 50% confidence ellipsoid in the case of a Gaussian distribution, which is slightly different from the 50% tolerance ellipsoid presented in Section 3.2, in the sense that the constraint $(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \boldsymbol{P} (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \leq 1$ equals $D_M^2(\boldsymbol{\mu}, \bar{\boldsymbol{\mu}}) \leq M$ and $M \approx \chi^2_{M,0.5}$. The parameters ρ_x and ρ_y are taken to be the maximum deviations between the covariances and the nominal covariances in the Frobenius norm sense over the set of resamples.

The paper also shows that for a specific type of uncertainty model, i.e. the product form uncertainty model $\mathcal{U} = \mathcal{M} \times \mathcal{S}$, where $\mathcal{M}$ is the set of possible means and $\mathcal{S}$ is the set of possible covariances, another equal optimization problem exists that produces the same results as (5) and is less computationally expensive. For this model, (5) can be written as

$$\begin{aligned} \text{minimize} \quad & (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T \left(\max_{(\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y) \in \mathcal{S}} \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y \right)^{-1} (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y) \\ \text{subject to} \quad & (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y) \in \mathcal{M}. \end{aligned}$$

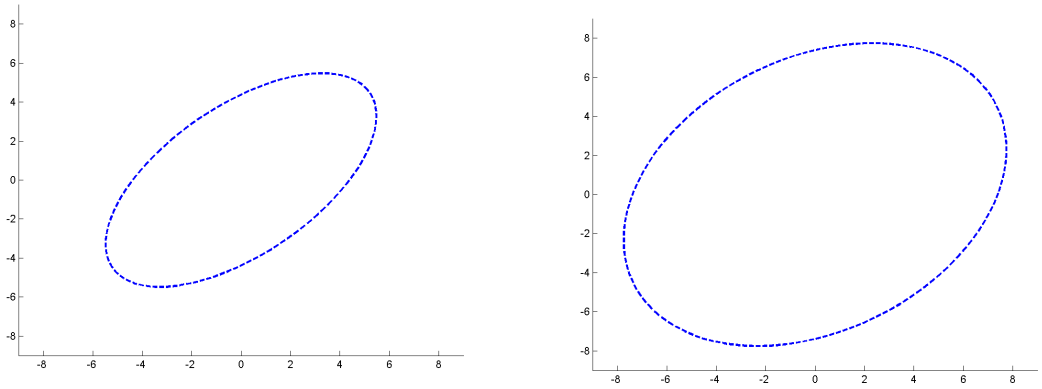
We find that $\max_{(\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y) \in \mathcal{S}} \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y = \bar{\boldsymbol{\Sigma}}_x + \bar{\boldsymbol{\Sigma}}_y + (\rho_x + \rho_y) \boldsymbol{I}$ (see, *e.g.*, [21]) with $\boldsymbol{I}$ as the identity matrix and therefore (5) equals

$$\begin{aligned} \text{minimize} \quad & (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T (\bar{\boldsymbol{\Sigma}}_x + \bar{\boldsymbol{\Sigma}}_y + (\rho_x + \rho_y) \boldsymbol{I})^{-1} (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y) \\ \text{subject to} \quad & (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y) \in \mathcal{M}, \end{aligned}$$

of which the outcomes $\check{\boldsymbol{\mu}}_x$ and $\check{\boldsymbol{\mu}}_y$ are used to compute the robust discriminant

$$\boldsymbol{w} = (\bar{\boldsymbol{\Sigma}}_x + \bar{\boldsymbol{\Sigma}}_y + (\rho_x + \rho_y)\boldsymbol{I})^{-1}(\check{\boldsymbol{\mu}}_x - \check{\boldsymbol{\mu}}_y).$$

Since the nominal covariances $\bar{\boldsymbol{\Sigma}}_x$ and $\bar{\boldsymbol{\Sigma}}_y$ are closely related to the sample estimates of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$, we can see that $\check{\boldsymbol{\Sigma}}_x = \bar{\boldsymbol{\Sigma}}_x + \rho_x \boldsymbol{I}$ and $\check{\boldsymbol{\Sigma}}_y = \bar{\boldsymbol{\Sigma}}_y + \rho_y \boldsymbol{I}$ are reshaped sample covariances: they now have a greater variance in the individual dimensions while the covariance of the dimensions remain the same, i.e. the covariances have become relatively smaller than the variances. Visually, this creates broader covariances, see Figure 3.



(a) Visualization of $\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$

(b) Visualization of $\boldsymbol{\Sigma} + 3\boldsymbol{I}$

Figure 3: The influence of a relatively smaller covariance.

This worst-case estimation of the covariances leads one to believe that it is only useful for a small sample size: given a data set with a small sample size, one should expect that, if we were to take more samples from the same distribution, there is a probability that these samples will lie wider, inducing a higher variance and lower coviariance. For this type of situation, this estimator would be appropriate. For a situation where outliers already influence the nominal covariance, it is not appropriate. Therefore, the worst-case estimates will probably not be effective as a robust method against outliers for classification. In Figure 4 we see visualizations of the sample covariance and the worst-case covariance, based on the same sample set given in Section 2.3. As expected, the sample covariance ellipsoids are completely encompassed by the worst-case covariance ellipsoids.

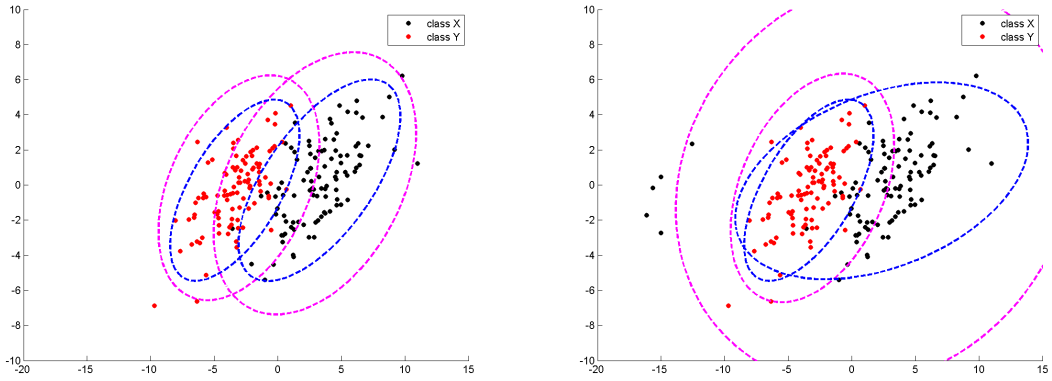
3.2 The $t_{M,p}$ estimates

In this section we will see that, if we use the linear transformation $\boldsymbol{V}$ for the multivariate normal distribution $\boldsymbol{X}$ as in (9), we obtain the equality

$$D_M^2(\boldsymbol{x}) \overset{(11)}{=} D_M^2(\boldsymbol{V}^T\boldsymbol{x}) \overset{(10)}{=} \sum_{i=1}^{M} \left(\frac{x_i' - \mu_i'}{\lambda_i} \right), \tag{6}$$

from which we can obtain a tolerance ellipsoid that theoretically encompasses a fraction p of n samples if $\lim_{n\to\infty}$, defined by the set of points,

$$\left\{ \boldsymbol{t}_{M,p} \in \mathbb{R}^{M\times 1} \,\middle|\, D_M^2(\boldsymbol{t}) \overset{(8)}{=} \chi_{M,p}^2 \right\}. \tag{7}$$



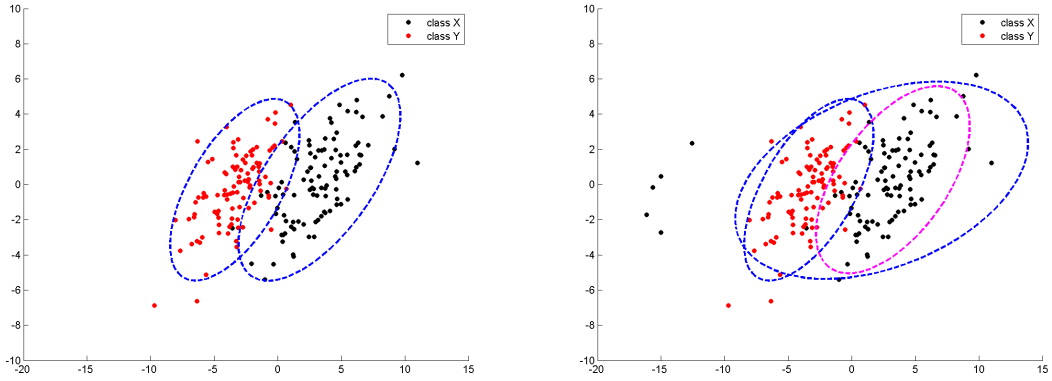
(a) Without outliers. (b) With outliers.

Figure 4: The blue ellipsoids represent the covariance ellipsoids, the magenta ellipsoids represent the worst-case covariance ellipsoids.

The equalities in Equations (6) and (7) are derived in Sections 3.2.1 to 3.2.3.

Assuming that outliers are samples that are distanced furthest from our mean in the D_M sense (see Section 3.2.3) and make up a fraction $1-p$ of our available M-dimensional data, we can rid them from our data set by removing all samples from our data that fall outside of our tolerance ellipsoid. The tolerance ellipsoid is defined per distribution by the set of points $\boldsymbol{t}_{M,p}$ where the estimating process of the means and covariances included the outliers. Then, we can re-estimate our mean and covariance with (almost) all outliers excluded. These re-estimates will be called *the* $\boldsymbol{t}_{M,p}$ *estimates*, which should not be confused with general $\boldsymbol{t}_{M,p}$ tolerance ellipsoids.

However, one must obtain an idea/estimate of this fraction of outliers and assume that the data is multivariate normally distributed. When the estimation of the fraction $1-p$ is too large, one might delete non-outliers.

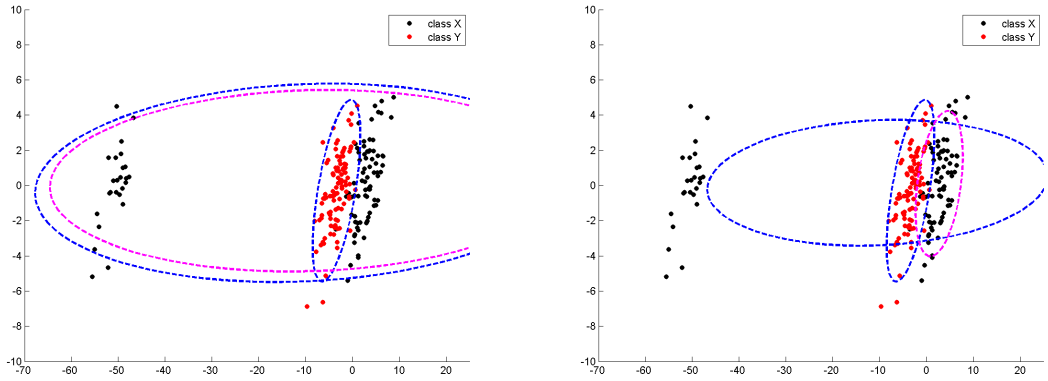


(a) Scatter plot of the sample set. (b) Scatter plot of the sample set with outliers.

Figure 5: Example of two 2-dimensional classes.

An example of the $\boldsymbol{t}_{2,0.95}$ ellipsoid is displayed in Figure 5 as the blue ellipsoids, based on the same sample set given in Section 2.3. The $\boldsymbol{t}_{M,0.95}$ ellipsoid can be regarded as a sample covariance ellipsoid,

since it shows where the two standard deviations boundary lies. Again, we see what the influence of outliers are on the sample covariance. The magenta ellipsoid represents the $\boldsymbol{t}_{M,0.95}$ estimate. It seems to be an accurate representation of the true covariance. In this case, using the tolerance ellipsoid to obtain the $\boldsymbol{t}_{M,p}$ estimates is a robust method.



(a) $\boldsymbol{t}_{M,0.95}$ is too large.

(b) $\boldsymbol{t}_{M,0.7}$ is more appropriate.

Figure 6: The blue ellipsoids represent the covariance ellipsoids, the magenta ellipsoids represent the worst-case covariance ellipsoids.

However, if the mean of the outliers is distanced further from the mean of the class and the number of outliers increases, we see that the $\boldsymbol{t}_{2,0.95}$ tolerance ellipsoid also encompasses many outliers, see Figure 6a. In this example, the number of outliers is a fraction 0.25 of the total sample set of class X. The $\boldsymbol{t}_{2,0.95}$ estimates are not accurate estimates. If the fraction of samples we want to exclude with the tolerance ellipsoid is adjusted to 0.3, which is slightly above the fraction of outliers, Figure 6b shows that some real samples of class X will be excluded as well. The blue line represents the $\boldsymbol{t}_{2,0.7}$ tolerance ellipsoid. If then we compute the $\boldsymbol{t}_{2,0.7}$ estimates, the covariance will be smaller than it should be, which is apparent by looking at the magenta ellipsoid in Figure 6b.

3.2.1 The chi-squared distribution

The chi-squared distribution with M degrees of freedom, χ^2_M, is a distribution of a sum of squares of M independent standard normally distributed random variables:

$$\text{if } X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \text{ are independent, then } Y = \sum_{i=1}^{M} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2_M.$$

Since Y is a sum of squares, we have that $|Y| = Y$. Therefore, $P[Y \leq y] = p$ indicates that the probability that a sample taken from Y falls within the interval $[-y, y]$ is p. To easily find y given p, we define the quantile function (inverse cumulative distribution function) of the chi-squared distribution as follows: if we let $F_{\chi^2_M}(y) = P[Y \leq y]$ be the cumulative density function of χ^2_M, then

$$F_{\chi^2_M}(y) = p \quad \Longleftrightarrow \quad \chi^2_{M,p} := y.$$

Fortunately, $\chi^2_{M,p}$ is given in MATLAB as the function `chi2inv(p,M)`. Now, given p, we can find the corresponding interval $[-y, y]$, such that

$$P\left[\sum_{i=1}^{M} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \leq \chi^2_{M,p} \right] = p,$$

for independent random variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Therefore, the tolerance ellipsoid that encompasses approximately a fraction p of our samples is defined by the set of points,

$$\left\{ \boldsymbol{t}_{M,p} \in \mathbb{R}^{M\times 1} \;\middle|\; \sum_{i=1}^{M} \left(\frac{t_i - \mu_i}{\sigma_i} \right)^2 = \chi^2_{M,p} \right\}. \tag{8}$$

3.2.2 Obtaining independent normal random variables

A set of $i = 1, \ldots, M$ random variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ can be expressed as a multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is an $M \times M$ diagonal matrix, then these random variables X_i are independent.

Given *dependent* random variables X_i, we want to transform the multivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is *not* a diagonal matrix, to a multivariate normal distribution $\boldsymbol{X}' \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, where $\boldsymbol{\Sigma}'$ *is* a diagonal matrix. Therefore, we compute the eigendecomposition of the matrix $\boldsymbol{\Sigma}$: a diagonalizable matrix $\boldsymbol{A}$ can be factorized into its eigenvalues and eigenvectors. If, in addition, $\boldsymbol{A}$ is positive-semidefinite, it can always be expressed as $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ are the normalized eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues [22]. Since a covariance matrix is always positive-semidefinite, we get, by projecting $\boldsymbol{X}$ onto $\boldsymbol{V}$, according to Theorem 1,

$$\begin{aligned} \boldsymbol{X}' &= \boldsymbol{V}^T\boldsymbol{X}, \\ \boldsymbol{\mu}' &= \boldsymbol{V}^T\boldsymbol{\mu}, \\ \boldsymbol{\Sigma}' &= \boldsymbol{V}^T\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{\Lambda}, \end{aligned} \tag{9}$$

and we have the desired linear transformation of $\boldsymbol{X}$ where the covariance matrix is diagonal.

3.2.3 Mahalanobis distance

The Mahalanobis Distance $D_M(\boldsymbol{x}, \boldsymbol{y})$ is a multi-dimensional generalization for measuring how many standard deviations away a sample $\boldsymbol{x}$ is from another sample $\boldsymbol{y}$ of the same distribution [16]. Let us define $D_M(\boldsymbol{x}) = D_M(\boldsymbol{x}, \boldsymbol{\mu})$. Given the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, the square of the relative distance $D_M(\boldsymbol{x})$ is given by

$$D_M^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}).$$

If $\boldsymbol{\Sigma}$ is a diagonal matrix, it is easily shown that

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{M} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2, \tag{10}$$

where x_i, μ_i and σ_i are the i^{th} elements of $\boldsymbol{x}, \boldsymbol{\mu}$ and the diagonal of $\boldsymbol{\Sigma}$, respectively. If we transform $\boldsymbol{x}$ according to (9) and use the fact that $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ implies that $\boldsymbol{V}^{-1} = \boldsymbol{V}^T$, from which we obtain the identity

$$\left(\boldsymbol{V}^T\boldsymbol{\Sigma}\boldsymbol{V}\right)^{-1} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{V}^T\right)^{-1} = \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V},$$

we can achieve the equality

$$D_M^2(\boldsymbol{V}^T\boldsymbol{x}) = D_M^2(\boldsymbol{x}). \tag{11}$$

Notice that this equality only holds for linear transformations using a unitary matrix such as $\boldsymbol{V}$ in (9).

4 Numerical results

To examine the effect of outliers on classification performance using Fisher LDA, we compare the success rates of the Fisher discriminant based on the regular sample estimates, the worst-case estimates and the $\boldsymbol{t}_{M,p}$ estimates. A robust version would perform better than the regular version if the success rate of the robust version is higher.

There will be $2^3 = 8$ configurations: the success rates of the three versions of the Fisher discriminant is tested by varying between two values for every variable. These variables are the sample size, fraction of outliers and the outlier means. For every configuration 1.000 data sets will be generated for two classes X and Y, real samples will be replaced with outliers according to the outlier fraction and outlier means and the performance of the three versions of the Fisher discriminant is tested on these data sets. The success rate for every configuration will be the mean of the success rates of each of the 1.000 generated data sets. The values for the variables will be

1. Sample size per class
   - Small: 20.
   - Large: 200.
2. Fraction of outliers
   - Small: drawn from a $\mathcal{N}(0.05, 0.03)$ distribution.
   - Large: drawn from a $\mathcal{N}(0.25, 0.03)$ distribution.
3. Distance outlier means from class means
   - Small: the means of the classes X and Y increased with ± 5, where $\pm$ indicates either 1 or -1 randomly.
   - Large: the means of the classes X and Y increased with ± 20 for every dimension.

The samples of the two classes will be drawn from multivariate normal distributions with means and covariances

$$\boldsymbol{\mu}_x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\mu}_y = \begin{bmatrix} -2 \\ 0 \end{bmatrix},$$
$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_x.$$

Outliers will be defined as a cluster of points not belonging to either X or Y. The outlier covariances for both X and Y will be the identity matrix. To include overestimation of the fraction of outliers, the fraction of outliers will be normally distributed with a variance of 0.03. The $\boldsymbol{t}_{M,0.95}$ estimates will be defined to remove 0.05 of samples above the mean fraction of outliers, i.e. the $\boldsymbol{t}_{2,0.9}$ and $\boldsymbol{t}_{2,0.7}$ estimates will be employed for the small and large fraction of outliers, respectively.

The expectations are that the worst-case estimates might perform better with a small sample size, although the outliers will interfere with its performance. The $\boldsymbol{t}_{M,0.95}$ estimates work best with a small distance of the outlier means, however, as we have seen, the configuration with large values for outlier fraction and distance might produce very poor results.

The success rate of classification without outliers and a large sample size is 0.87, which will be the reference success rate, i.e. we cannot expect the success rates of classification with outliers to be higher than this reference success rate. However, we do want any version of the Fisher discriminant to produce success rates close to the reference success rate. The results of the experiment are shown in Table 1. A configuration is indicated by a combination of S's and L's, where S and L indicate small and large values, respectively. They are in the order of the variables as given above.

For all three versions of the Fisher discriminant, we see that increasing the outlier fraction and mean distance negatively influences the success rates, while the configuration with small outlier fraction and

Table 1: Success rates of three versions of the Fisher discriminant.

	SSS	SSL	SLS	SLL	LSS	LSL	LLS	LLL
Regular estimates	0.836	0.697	0.758	0.641	0.856	0.729	0.786	0.660
Worst-case estimates	0.822	0.638	0.685	0.619	0.851	0.731	0.786	0.629
$\boldsymbol{t}_{M,p}$ estimates	0.846	0.848	0.751	0.735	0.866	0.867	0.787	0.810

mean distance does not differ much compared to the reference success rate. The worst-case estimates produce their worst results when the outlier distance is large. The fraction of outliers does not lead to a great difference compared to the regular sample estimates. In none of the configurations do the worst-case estimates perform best. The $\boldsymbol{t}_{M,p}$ estimates performed best in all but one configuration (SLS). However, when the outlier fraction is large and outlier distance is small, employing the $\boldsymbol{t}_{M,p}$ estimates is not more effective than the regular sample estimates.

5 Conclusion

In the case where outliers are defined as a group of points lying further from the mean in the MD sense, experiments were conducted that show the performance of the Fisher discriminant based on the regular sample estimates, the worst-case estimates and the $\boldsymbol{t}_{M,p}$ estimates. The best performance is given by the $\boldsymbol{t}_{M,p}$ estimates, whereas the worst-case estimates did not show better performance in any of the configurations.

The worst-case estimates take on a specific shape which is not desirable in case of outliers. It might be employed in case of small sample sizes, as [15] seems to indicate. However, this paper did not investigate its performance on small sample sizes without outliers. This suggests to investigate the performance of the worst-case estimates on small sample sizes in future research. Another suggestion for future research is to construct different constraints for the covariance matrices that are used in the optimization problem, since the intuitive idea seems plausible.

The $\boldsymbol{t}_{M,p}$ estimates are predictable to some extent and can be employed in many cases. However, there are some cases where the regular sample estimates seem to perform at least as well. Future research may wish to alter or finetune the computation of the $\boldsymbol{t}_{M,p}$ estimates so it can be used as a robust method in these cases as well.

The experiments in this paper were fully based on multivariate normally distributed data. Future research may wish to apply these methods to real world data or implement different definitions of outliers.

Appendix

Theorem 1 (Linear transformation of multivariate normal distribution)
Let $\boldsymbol{X}$ be an $M \times 1$ multivariate normal random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{A}$ be an $L \times 1$ real vector and $\boldsymbol{B}$ an $L \times M$ full-rank real matrix. Then the $L \times 1$ random vector $\boldsymbol{Y}$ defined by

$$\boldsymbol{Y} = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{X}$$

has a multivariate normal distribution with mean

$$E[\boldsymbol{Y}] = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{\mu}$$

and covariance matrix

$$Cov[\boldsymbol{Y}] = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T$$

PROOF The joint moment generating function of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left[e^{\boldsymbol{t}^T\boldsymbol{X}}\right] = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}$$

Therefore, the joint moment generating function of $\boldsymbol{Y}$ is

$$\begin{aligned}
M_{\boldsymbol{Y}}(\boldsymbol{t}) &= E\left[e^{\boldsymbol{t}^T(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{X})}\right] = E\left[e^{\boldsymbol{t}^T\boldsymbol{A}}e^{\boldsymbol{t}^T\boldsymbol{B}\boldsymbol{X}}\right] \\
&= e^{\boldsymbol{t}^T\boldsymbol{A}}E\left[e^{\boldsymbol{t}^T\boldsymbol{B}\boldsymbol{X}}\right] \quad \text{(because } e^{\boldsymbol{t}^T\boldsymbol{A}} \text{ is a scalar)} \\
&= e^{\boldsymbol{t}^T\boldsymbol{A}}M_{\boldsymbol{X}}\left(\boldsymbol{B}^T\boldsymbol{t}\right) \\
&= e^{\boldsymbol{t}^T\boldsymbol{A}}e^{\boldsymbol{t}^T\boldsymbol{B}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T\boldsymbol{t}} \\
&= e^{\boldsymbol{t}^T(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\mu})+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T\boldsymbol{t}}
\end{aligned}$$

which is the joint moment generating function of a multivariate normal distribution with mean $\boldsymbol{A}+\boldsymbol{B}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T$. Since two random vectors have the same distribution when they have the same joint moment generating function, $\boldsymbol{Y}$ has a multivariate normal distribution with mean $\boldsymbol{A} + \boldsymbol{B}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T$. □

Theorem 2 (Fisher Linear Discriminant Analysis)
The Fisher discriminant ratio is given by

$$f(\boldsymbol{w}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y) = \frac{\boldsymbol{w}^T(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T\boldsymbol{w}}{\boldsymbol{w}^T(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)\boldsymbol{w}} = \frac{(\boldsymbol{w}^T(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y))^2}{\boldsymbol{w}^T(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)\boldsymbol{w}}, \tag{12}$$

A discriminant that maximizes the Fisher discriminant ratio is given by

$$\boldsymbol{w} = (\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)$$

which gives the maximum Fisher discriminant ratio

$$\max_{\boldsymbol{w}\neq\boldsymbol{0}} f(\boldsymbol{w}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y) = (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)$$

PROOF The proof is given in [15]. □

References

[1] Bishop, Christopher M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Härdle, Wolfgang and Simar, Léopold. *Applied multivariate statistical analysis*, volume 22007. Springer, 2007.

[3] Xu, Huan and Caramanis, Constantine and Sanghavi, Sujay. Robust PCA via Outlier Pursuit. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.

[4] Zhang, Huishuai and Zhou, Yi and Liang, Yingbin. Analysis of robust PCA via local incoherence. In *Advances in Neural Information Processing Systems*, pages 1819–1827, 2015.

[5] Candès, Emmanuel J. and Li, Xiaodong and Ma, Yi and Wright, John. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[6] Yi, Juneho and Yang, Heesung and Kim, Yuho. Enhanced Fisherfaces for Robust Face Recognition. In Lee, Seong-Whan and Bülthoff, Heinrich H. and Poggio, Tomaso, editor, *Biologically Motivated Computer Vision*, pages 502–511, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[7] Liu, Chengjun and Wechsler, Harry. Enhanced fisher linear discriminant models for face recognition. In *Proceedings. Fourteenth International Conference on Pattern Recognition*, volume 2, pages 1368–1372. IEEE, 1998.

[8] Lu, Juwei and Plataniotis, Konstantinos N. and Venetsanopoulos, Anastasios N. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*, 26(2):181–191, 2005.

[9] Friedman, Jerome H. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[10] Gnanadesikan, Ramanathan and Kettenring, John R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124, 1972.

[11] Ronald A. Fisher. The use of multiple measurements in taxonomix problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[12] Welling, Max. Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 3(1), 2005.

[13] Balakrishnama, Suresh and Ganapathiraju, Aravind. Linear Discriminant Analysis - A Brief Tutorial. 11:, 01 1998.

[14] Tharwat, Alaa and Gaber, Tarek and Ibrahim, Abdelhameed and Hassanien, Aboul Ella. Linear discriminant analysis: A detailed tutorial. *Ai Communications*, 30:169–190,, 05 2017.

[15] Seung-Jean Kim and Alessandro Magnani and Stephen P. Boyd. Robust Fisher Discriminant Analysis. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 659–666, 2005.

[16] Hubert, Mia and Debruyne, Michiel and Rousseeuw, Peter J. . Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2009.

[17] Rousseeuw, Peter J. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

[18] Gupta, Shanti S. Probability integrals of multivariate normal and multivariate t^1. *The Annals of mathematical statistics*, pages 792–828, 1963.

[19] Chew, Victor. Confidence, prediction, and tolerance regions for the multivariate normal distribution. *Journal of the American Statistical Association*, 61(315):605–617, 1966.

[20] Efron, Bradley and Tibshirani, Robert J. *An introduction to the bootstrap*. CRC press, 1994.

[21] Lanckriet, Gert R.G. and Ghaoui, Laurent El and Bhattacharyya, Chiranjib and Jordan, Michael I. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3(Dec):555–582, 2002.

[22] Abdi, Hervé. The eigen-decomposition: Eigenvalues and eigenvectors.