Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data

Robert J. Durrant
School of Computer Science
University of Birmingham
Edgbaston, UK, B15 2TT
r.j.durrant@cs.bham.ac.uk

Ata Kabán
School of Computer Science
University of Birmingham
Edgbaston, UK, B15 2TT
a.kaban@cs.bham.ac.uk

ABSTRACT

We consider random projections in conjunction with classification, specifically the analysis of Fisher's Linear Discriminant (FLD) classifier in randomly projected data spaces.

Unlike previous analyses of other classifiers in this setting, we avoid the unnatural effects that arise when one insists that all pairwise distances are approximately preserved under projection. We impose no sparsity or underlying low-dimensional structure constraints on the data; we instead take advantage of the class structure inherent in the problem. We obtain a reasonably tight upper bound on the estimated misclassification error on average over the random choice of the projection, which, in contrast to early distance preserving approaches, tightens in a natural way as the number of training examples increases. It follows that, for good generalisation of FLD, the required projection dimension grows logarithmically with the number of classes. We also show that the error contribution of a covariance misspecification is always no worse in the low-dimensional space than in the initial high-dimensional space. We contrast our findings to previous related work, and discuss our insights.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Theory

Keywords

Dimensionality Reduction, Random Projection, Compressed Learning, Linear Discriminant Analysis, Classification

1. INTRODUCTION

Dimensionality reduction via random projections has attracted considerable interest for its theoretical guarantees to approximately preserve pairwise distances globally, and for the computational advantages that it offers. While this theory is quite well developed, much less is known about exact guarantees on the performance and behaviour of subsequent data analysis methods that work with the randomly projected data. Obtaining such results is not straightforward, and is subject to ongoing recent research efforts.

In this paper we consider the supervised learning problem of classifying a query point $\mathbf{x}_q \in \mathbb{R}^d$ as belonging to one of several Gaussian classes using Fisher's Linear Discriminant (FLD) and the classification error arising if, instead of learning the classifier in the data space $\mathbb{R}^d$, we instead learn it in some low dimensional random projection of the data space $R(\mathbb{R}^d) = \mathbb{R}^k$, where $R \in \mathcal{M}_{k\times d}$ is a random projection matrix with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, 1/d)$. FLD is one of the most enduring methods for data classification, yet it is simple enough to be tractable to detailed formal analysis in the projected domain.

The main practical motivations behind this research are the perspective of mitigating the issues associated with the curse of dimensionality by working in a lower dimensional space, and the possibility of not having to collect or store the data in its original high dimensional form.

A number of recent studies consider efficient learning in low dimensional spaces. For example, in [5] Calderbank et al demonstrate that if the high dimensional data points have a sparse representation in some linear basis, then it is possible to train a soft-margin SVM classifier on a low dimensional projection of that data whilst retaining a classification performance that is comparable to that achieved by working in the original data space. However, the data points must be capable of a sparse representation which for a general class-learning problem may not be the case.

In [9] Davenport et al prove high probability bounds (over the choice of random projection) on a series of signal processing techniques, among which are bounds on signal classification performance for a single test point using Neyman-Pearson detector and on the estimation of linear functions of signals from few measurements when the set of potential signals is known but with no sparsity requirement on the signal. Similarly, in [12] Haupt et al demonstrate that $(m + 1)$-ary hypothesis testing can be used to specify, from few measurements, to which of a known collection of prototypes a signal belongs. More recently bounds on least squares regression in projected space, with no sparsity requirement on the data,

KDD'10, July 25–28, 2010, Washington, DC, USA.

have been presented by Maillard and Munos in [14].

We also do not require our data to have a sparse representation. We ask whether we can still learn a classifier using a low-dimensional projection of the data (the answer is 'yes'), and to what dimensionality k can the data be projected so that the classifier performance is still maintained.

We approach these questions by bounding the probability that the classifier assigns the wrong class to a query point if the classifier is learned in the projected space. Such bounds on the classification error for FLD in the data space are already known, for example they are given in [4, 15], but in neither of these papers is classification error in the projected domain considered; indeed in [9] it is stated that establishing the probability of error for a classifier in the projected domain is, in general, a difficult problem.

As we shall see in the later sections, our bounds are reasonably tight and as a direct consequence we find that the projection dimension required for good generalisation of FLD in the projection space depends logarithmically on the number of classes. This is, of course, typically very low compared to the number of training examples. Unlike the bounds in [3], where the authors' use of the Johnson-Lindenstrauss Lemma[1] has the unfortunate side-effect that their bound loosens as the number of training examples increases, because of the increased accuracy of mean estimates our bound will tighten with more training data.

The structure of the remainder of the paper is as follows: We briefly describe the supervised learning problem and describe the FLD classifier. We then bound the probability that FLD will misclassify an unseen query point $\mathbf{x}_q$ having learned the classifier in a low dimensional projection of the data space and applied it to the projected query point $R\mathbf{x}_q$. This is equivalent to bounding the expected $(0,1)$-loss of the projected FLD classifier. Finally we discuss our findings and indicate some possible future directions for this work.

2. THE SUPERVISED LEARNING PROBLEM

In a supervised learning problem we observe N examples of training data $\mathcal{T}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $(\mathbf{x}_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}$ some (usually unknown) distribution with $\mathbf{x}_i \sim \mathcal{D}_x \subseteq \mathbb{R}^d$ and $y_i \sim \mathcal{C}$, where $\mathcal{C}$ is a finite collection of class labels partitioning $\mathcal{D}$.

For a given class of functions $\mathcal{H}$, our goal is to learn from $\mathcal{T}_N$ the function $\hat{h} \in \mathcal{H}$ with the lowest possible generalisation error in terms of some loss function $\mathcal{L}$. That is, find $\hat{h}$ such that $\mathcal{L}(\hat{h}) = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathbf{x}_q}[\mathcal{L}(h)]$, where $(\mathbf{x}_q, y) \sim \mathcal{D}$ is a query point with unknown label y.

Here we use the $(0,1)$-loss $\mathcal{L}_{(0,1)}$ as a measure of performance defined by:

$$\mathcal{L}_{(0,1)} = \begin{cases} 0 & \text{if } \hat{h}(\mathbf{x}_q) = y \\ 1 & \text{otherwise.} \end{cases}$$

In the case we consider here, the class of functions $\mathcal{H}$ consists of instantiations of FLD learned on randomly-projected data, $\mathcal{T}_N = \{(R\mathbf{x}_i, y_i) : R\mathbf{x} \in \mathbb{R}^k, \mathbf{x} \sim \mathcal{D}_x\}$ and we seek to bound the probability that an arbitrarily drawn and previously unseen query point $\mathbf{x}_q \sim \mathcal{D}_x$ is misclassified by the learned classifier. Our approach is to bound:

$$\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y : \mathbf{x}_q \sim \mathcal{D}_x] = \mathrm{E}_{\mathbf{x}_q}[\mathcal{L}_{(0,1)}(\hat{h}(\mathbf{x}_q), y)]$$

and then to bound the corresponding probability in a random projection of the data:

$$\Pr_{R,\mathbf{x}_q}[\hat{h}(R\mathbf{x}_q) \neq y : \mathbf{x}_q \sim \mathcal{D}_x] = \mathrm{E}_{R,\mathbf{x}_q}[\mathcal{L}_{(0,1)}(\hat{h}(R\mathbf{x}_q), y)]$$

For concreteness and tractability, we will do this for the Fisher Linear Discriminant (FLD) classifier, which is briefly reviewed below.

3. FISHER'S LINEAR DISCRIMINANT

FLD is a generative classifier that seeks to model, given training data $\mathcal{T}_N$, the optimal decision boundary between classes. It is a successful and widely used classification method. The classical version is formulated for 2-class problems, as follows. If π_0, $\Sigma = \Sigma_0 = \Sigma_1$ and μ_0 and μ_1 are known then the optimal classifier is given by Bayes' rule [4]:

$$h(\mathbf{x}_q) = \mathbf{1}\left\{\log\frac{(1-\pi_0)f_1(\mathbf{x}_q)}{\pi_0 f_0(\mathbf{x}_q)} > 0\right\}$$
$$= \mathbf{1}\left\{\log\frac{1-\pi_0}{\pi_0} + (\mu_1-\mu_0)^T\Sigma^{-1}\left(\mathbf{x}_q - \frac{(\mu_0+\mu_1)}{2}\right) > 0\right\}$$

where $\mathbf{1}(A)$ is the indicator function that returns one if A is true and zero otherwise, and f_y is the probability density function of the y-th data-class, in its simplest form the multivariate Gaussian $\mathcal{N}(\mu_y, \Sigma)$, namely:

$$\left((2\pi)^{d/2}\det(\Sigma)^{1/2}\right)^{-1}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma^{-1}(\mathbf{x}-\mu_y)\right)$$

In the sequel, we shall assume that the observations $\mathbf{x}$ are drawn from one of $m+1$ multivariate Gaussian classes[2] $\mathcal{D}_x = \sum_{y=0}^m \pi_y\mathcal{N}(\mu_y, \Sigma)$ with unknown parameters μ_y and Σ that we need to estimate from training data. As usual, μ_y is a vector of means and Σ is a full-rank covariance matrix.

4. RESULTS

4.1 Roadmap

Our main result is theorem 4.8, which bounds the estimated misclassification probability of the two-class Fisher's Linear Discriminant classifier (FLD) when the classifier is learnt from a random projection of the original training data and the class label is assigned to a query point under the same random projection of the data. Our bound holds on average over the random projection matrices R, in contrast to other bounds in the literature [8, 3] where the techniques depend upon the fact that under suitable conditions randomly projected data satisfies the Johnson Lindenstrauss Lemma with high probability.

4.1.1 Structure of proof

We commence by an analysis of the error probability of FLD in the data space, which provides us with an upper bound on the error probability in a suitable form to make our subsequent average error analysis tractable in the projection space. Here we also show how we can deal with multiclass cases, and we highlight the contribution of the

[1] For a proof of the JLL and its application to random projection matrices see e.g. [8, 1]

[2] The Gaussianity of the true class distributions can be relaxed to sub-Gaussian class densities without any effort — see comment in the Appendix section 7.0.4.

Table 1: Notation used throughout this paper

Random vector	$\mathbf{x}$
Observation/class label pair	$(\mathbf{x}_i, y_i)$
Query point (unlabelled observation)	$\mathbf{x}_q$
Set of $m+1$ class labels partitioning the data	$\mathcal{C} = \{0, 1, \dots, m\}$
Random projection matrix	R
'Data space' - real vector space of d dimensions	$\mathbb{R}^d$
'Projected space' - real vector space of $k \leqslant d$ dimensions	$\mathbb{R}^k$
Mean of class $y \in \mathcal{C}$	μ_y
Sample mean of class $y \in \mathcal{C}$	$\hat{\mu}_y$
Covariance matrix of the Gaussian distribution $\mathcal{D}_{x_y}$	Σ
Estimated (model) covariance matrix of the Gaussian distribution $\mathcal{D}_{x_y}$	$\hat{\Sigma}$
Prior probability of membership of class $y \in \mathcal{C}$	π_y
Estimated prior probability of membership of class $y \in \mathcal{C}$	$\hat{\pi}_y$

estimated error and the estimation error terms in the overall generalisation error. Our data space analysis also has the advantage of being slightly more general than existing ones in e.g. [4, 15], in that it holds also for non-Gaussian but sub-Gaussian classes.

We then review some tools from matrix analysis in preparation for the main section. The proof of our main result, namely the bound on the estimated error probability of FLD in a random projection of the data space, then follows, along with the required dimensionality of the projected space as its direct consequence.

4.2 Preliminaries: Data space analysis of FLD

4.2.1 Bound on two-class FLD in the data space

Our starting point is the following lemma which, for completeness, is proved in the appendix:

Lemma 4.1. *Let $\mathbf{x_q} \sim \mathcal{D}_x$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Assume that we have sufficient data so that $\kappa_y = (\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1} (\hat{\mu}_{\neg y} - \hat{\mu}_y) - 2 \log \frac{1-\hat{\pi}_y}{\hat{\pi}_y} > 0$ (i.e. μ_y and $\hat{\mu}_y$ lie on the same side of the estimated hyperplane) $\forall y, \neg y \in \mathcal{C} = \{0, 1\}, y \neq \neg y$. Then the probability that $\mathbf{x}_q$ is misclassified is bounded above by $Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] \leqslant$*

$$\pi_0 \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) - 2 \log \frac{1-\hat{\pi}_0}{\hat{\pi}_0} \right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} \right) +$$

$$(1-\pi_0) \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_1)^T \hat{\Sigma}^{-1} (\hat{\mu}_0 - \hat{\mu}_1) - 2 \log \frac{\hat{\pi}_0}{1-\hat{\pi}_0} \right]^2}{(\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\mu}_0 - \hat{\mu}_1)} \right)$$

with μ_y the mean of the class from which $\mathbf{x}_q$ was drawn, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$, model covariance $\hat{\Sigma}$, true class priors π_0 and $1-\pi_0$, and estimated class priors $\hat{\pi}_0$ and $1-\hat{\pi}_0$.

4.2.2 Multi-class case

The multi-class version of FLD may be analysed in extension to the above analysis as follows:

Lemma 4.2. *Let $\mathcal{C} = \{0, 1, \dots, m\}$ be a collection of $m+1$ classes partitioning the data.*
Let $\mathbf{x_q} \sim \mathcal{D}_x$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Then, the probability that an unseen query point $\mathbf{x}_q$ is misclassified by FLD is given by $Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] \leqslant$:

$$\sum_{y=0}^{m} \pi_y \sum_{i \neq y}^{m} \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_y - \hat{\mu}_i - 2\mu_y)^T \hat{\Sigma}^{-1} (\hat{\mu}_y - \hat{\mu}_i) - 2 \log \frac{\hat{\pi}_i}{\hat{\pi}_y} \right]^2}{(\hat{\mu}_y - \hat{\mu}_i)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\mu}_y - \hat{\mu}_i)} \right) \tag{4.1}$$

Proof. The decision rule for FLD in the multi-class case is given by:

$$\hat{h}(\mathbf{x}_q) = j \iff j = \arg\max_i \{\Pr_y(i|\mathbf{x}_q)\} \quad y, j, i \in \mathcal{C}$$

Without loss of generality, we again take the correct class to be class 0 and we assume uniform estimated priors, the non-uniform case being a straightforward extension of lemma 4.1. Hence:

$$\hat{h}(\mathbf{x}_q) = 0 \iff \bigwedge_{i \neq 0} \{\Pr_y(0|\mathbf{x}_q) \geqslant \Pr_y(i|\mathbf{x}_q)\} \tag{4.2}$$

$$\iff \bigwedge_{i \neq 0} \left\{ \frac{\Pr_y(0|\mathbf{x}_q)}{\Pr_y(i|\mathbf{x}_q)} \geqslant 1 \right\} \tag{4.3}$$

and so misclassification occurs when:

$$\hat{h}(\mathbf{x}_q) \neq 0 \iff \bigvee_{i \neq 0} \left\{ \frac{\Pr_y(i|\mathbf{x}_q)}{\Pr_y(0|\mathbf{x}_q)} > 1 \right\}$$

Then since if $A \iff B$, $\Pr(A) = \Pr(B)$, we have:

$$\begin{aligned} \Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0] &= \Pr_{\mathbf{x}_q}\left[\bigvee_{i \neq 0} \left\{ \frac{\Pr_y(i|\mathbf{x}_q)}{\Pr_y(0|\mathbf{x}_q)} > 1 \right\} \right] \\ &\leqslant \sum_{i=1}^{m} \Pr_{\mathbf{x}_q} \left\{ \frac{\Pr_y(i|\mathbf{x}_q)}{\Pr_y(0|\mathbf{x}_q)} > 1 \right\} \qquad (4.4) \\ &= \sum_{i=1}^{m} \Pr_{\mathbf{x}_q} \left\{ \log \frac{\Pr_y(i|\mathbf{x}_q)}{\Pr_y(0|\mathbf{x}_q)} > 0 \right\} \qquad (4.5) \end{aligned}$$

where (4.4) follows by the union bound. Writing out (4.5) via Bayes' rule, we find a sum of 2-class error probabilities of the form that we have dealt with earlier, so (4.5) equals:

$$\sum_{i=1}^{m} \Pr_{\mathbf{x}_q} \left\{ \log \frac{1-\pi_0}{\pi_0} + (\hat{\mu}_i - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_i}{2} \right) > 0 \right\} \tag{4.6}$$

The result for $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0]$ given $y = 0$ now follows by applying the bounding technique used for the two-class case

m times to each of the m possible incorrect classes. The line of thought is then the same for $y = 1, \dots, y = m$ in turn.

Owing to the straightforward way in which the multiclass error bound boils down to sums of 2-class errors, as shown in lemma 4.2 above, it is therefore sufficient for the remainder of the analysis to be performed for the 2-class case, and for $m + 1$ classes the error will always be upper bounded by m times the greatest of the 2-class errors. This will be used later in Section 4.5.

Next, we shall decompose the FLD bound of lemma 4.1 into two terms, one of which will go to zero as the number of training examples increases. This gives us the opportunity to assess the contribution of these two sources of error separately.

4.2.3 Decomposition of data space bound as sum of estimated error and estimation error

Lemma 4.3. *Let $\mathbf{x_q} \sim \mathcal{D}_x$ and let $\mathcal{H}$ be the class of FLD functions and $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Write for the estimated error $\hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma, \hat{\pi}_0) =$*

$$\sum_{y=0}^{1} \pi_y \exp\left(-\frac{1}{8} \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) - 2 \log \frac{1-\hat{\pi}_y}{\hat{\pi}_y} \right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} \right) \tag{4.7}$$

and $B(\hat{\mu}_{0,1}, \hat{\mu}_{0,1}, \mu_y, \hat{\Sigma}, \Sigma, \hat{\pi}_0)$ for the right hand side of lemma 4.1. Then $Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] \leqslant$

$$\hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma, \hat{\pi}_y) + C \cdot \sum_{y,i} |\hat{\mu}_{yi} - \mu_{yi}| \tag{4.8}$$

with $C = \max_{y,i} \sup \left\{ \left| \frac{\partial B_{y \in \mathcal{C}}}{\partial \mu_{yi}} \right| \right\}$ a constant, μ_y the mean of the class from which $\mathbf{x}_q$ was drawn, estimated class means $\hat{\mu}_y$ with $\hat{\mu}_{yi}$ the i-th component, model covariance $\hat{\Sigma}$, and estimated class priors $\hat{\pi}_y$ and $1 - \hat{\pi}_y$.

Proof. We will use the mean value theorem[3], so we start by differentiating $B_{y \in \mathcal{C}}$ with respect to μ_0. Writing $\hat{B}_0$ and $\hat{B}_1$ for the two exp terms in (4.7), we have:

$$\nabla_{\mu_0} B = \pi_0 \hat{B}_0 \times \frac{1}{2} \kappa_0 \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \tag{4.9}$$

Since the exponential term is bounded between zero and one, the supremum of the i-th component of this gradient exists provided that $|\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0| < \infty$ and $|\hat{\mu}_1 - \hat{\mu}_0| < \infty$. So we have that

$$B \leqslant \pi_0 \hat{B}_0 + \max_i \sup \left\{ \left| \frac{\partial B_{y \in \mathcal{C}}}{\partial \mu_{0i}} \right| \right\} \sum_i |\hat{\mu}_{0i} - \mu_{0i}| \dots$$
$$\dots + (1 - \pi_0) \mathrm{Pr}_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 1 | y = 1] \tag{4.10}$$

Now applying the mean value theorem again w.r.t. μ_1 decomposes the latter term similarly, then taking the maximum over both classes yields the desired result. We call the two terms obtained in (4.8) the 'estimated error' and 'estimation error' respectively. The estimation error can be bounded using Chernoff bounding techniques, and converges to zero with increasing number of training examples.

In the remainder of the paper, we will take uniform model priors for convenience. Then the exponential terms of the estimated error (4.7) are all equal and independent of y, so using that $\sum_{y \in \mathcal{C}} \pi_y = 1$ the expression (4.7) of the estimated error simplifies to: $\hat{B}(\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma) =$

$$\exp\left(-\frac{1}{8} \cdot \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0) \right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)} \right) \tag{4.11}$$

We now have the groundwork necessary to prove our main result, namely a bound on this estimated misclassification probability if we choose to work with a k-dimensional random projection of the original data. From the results of lemma 4.3 and lemma 4.2, in order to study the behaviour of our bound, we may restrict our attention to the two-class case and we focus on bounding the estimated error term — which, provided sufficient training data, is the main source of error. Before proceeding, the next section gives some technical tools that will be needed.

4.3 Tools from matrix analysis

First, we note that due to linearity of both the expectation operator E$[\cdot]$, and the random projection matrix R, the projection of the true mean μ and sample mean $\hat{\mu}$ from the data space to the projected space coincides with the true mean $R\mu$ and the sample mean $R\hat{\mu}$ in the projected space. Furthermore, the projected counterparts of the true covariance matrix Σ and the model covariance $\hat{\Sigma}$ are given by $R\Sigma R^T$ and $R\hat{\Sigma}R^T$ respectively. Hence, we may talk about projected means and covariances unambiguously.

We will, in the proof that follows, make frequent use of several results. Apart from lemma 4.6 these are more or less well-known, but for convenience we state them here for later reference.

Lemma 4.4 (Rayleigh quotient. ([13], Theorem 4.2.2 Pg 176))**.** *If Q is a real symmetric matrix then its eigenvalues λ satisfy:*

$$\lambda_{\min}(Q) \leqslant \frac{\mathbf{v}^T Q \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \leqslant \lambda_{\max}(Q)$$

and, in particular:

$$\lambda_{\min}(Q) = \min_{\mathbf{v} \neq 0} \frac{\mathbf{v}^T Q \mathbf{v}}{\mathbf{v}^T \mathbf{v}} = \min_{\mathbf{v}^T \mathbf{v} = 1} \mathbf{v}^T Q \mathbf{v} \text{ and} \tag{4.12}$$

$$\lambda_{\max}(Q) = \max_{\mathbf{v} \neq 0} \frac{\mathbf{v}^T Q \mathbf{v}}{\mathbf{v}^T \mathbf{v}} = \max_{\mathbf{v}^T \mathbf{v} = 1} \mathbf{v}^T Q \mathbf{v} \tag{4.13}$$

Lemma 4.5 (Poincaré Separation Theorem. ([13], Corollary 4.3.16 Pg 190))**.** *Let S be a symmetric matrix $S \in \mathcal{M}_d$, let k be an integer, $1 \leqslant k \leqslant d$, and let $\mathbf{r}_1, \dots, \mathbf{r}_k \in \mathbb{R}^d$ be k orthonormal vectors. Let $T = \mathbf{r}_i^T S \mathbf{r}_j \in \mathcal{M}_k$ (that is, in our setting, the $\mathbf{r}_i^T$ are the rows and the $\mathbf{r}_j$ the columns of the random projection matrix $R \in \mathcal{M}_{k \times d}$ and so $T = RSR^T$). Arrange the eigenvalues λ_i of S and T in increasing magnitude, then:*

$$\lambda_i(S) \leqslant \lambda_i(T) \leqslant \lambda_{i+n-k}(S), \quad i \in \{1, \dots, k\}$$

and, in particular:

$$\lambda_{\min}(S) \leqslant \lambda_{\min}(T) \text{ and } \lambda_{\max}(T) \leqslant \lambda_{\max}(S)$$

[3] Mean value theorem in several variables: Let f be differentiable on S, an open subset of $\mathbb{R}^d$, let $\mathbf{x}$ and $\mathbf{y}$ be points in S such that the line between $\mathbf{x}$ and $\mathbf{y}$ also lies in S. Then: $f(\mathbf{y}) - f(\mathbf{x}) = (\nabla f((1-t)\mathbf{x} + t\mathbf{y}))^T (\mathbf{y} - \mathbf{x})$, $t \in (0, 1)$

Lemma 4.6 (Corollary to lemmata 4.4 and 4.5)**.** *Let Q be symmetric positive definite, such that $\lambda_{\min}(Q) > 0$ and so Q is invertible. Let $\mathbf{u} = R\mathbf{v}$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{u} \neq 0 \in \mathbb{R}^k$. Then:*

$$\mathbf{u}^T \left[RQR^T\right]^{-1} \mathbf{u} \geqslant \lambda_{\min}(Q^{-1})\mathbf{u}^T\mathbf{u} > 0$$

Proof: We use the eigenvalue identity $\lambda_{\min}(Q^{-1}) = 1/\lambda_{\max}(Q)$. Combining this identity with lemma 4.4 and lemma 4.5 we have:

$$\lambda_{\min}([RQR^T]^{-1}) = 1/\lambda_{\max}(RQR^T)$$

Since RQR^T is symmetric positive definite. Then by positive definiteness and lemma 4.5 it follows that:

$$\begin{aligned} & 0 < \lambda_{\max}(RQR^T) \leqslant \lambda_{\max}(Q) && (4.14)\\ \iff \quad & 1/\lambda_{\max}(RQR^T) \geqslant 1/\lambda_{\max}(Q) > 0 && (4.15)\\ \iff \quad & \lambda_{\min}([RQR^T]^{-1}) \geqslant \lambda_{\min}(Q^{-1}) > 0 && (4.16)\end{aligned}$$

And so by lemma 4.4:

$$\begin{aligned} \mathbf{u}^T \left[RQR^T\right]^{-1} \mathbf{u} \quad & \geqslant \quad \lambda_{\min}([RQR^T]^{-1})\mathbf{u}^T\mathbf{u} && (4.17)\\ & \geqslant \quad \lambda_{\min}(Q^{-1})\mathbf{u}^T\mathbf{u} && (4.18)\\ & = \quad \mathbf{u}^T\mathbf{u}/\lambda_{\max}(Q) > 0 && (4.19)\end{aligned}$$

Lemma 4.7 (Kantorovich Inequality. ([13], Theorem 7.4.41 Pg 444))**.** *Let Q be a symmetric positive definite matrix $Q \in \mathcal{M}_d$ with eigenvalues $0 < \lambda_{\min} \leqslant \ldots \leqslant \lambda_{\max}$. Then, for all $\mathbf{v} \in \mathbb{R}^d$:*

$$\frac{(\mathbf{v}^T\mathbf{v})^2}{(\mathbf{v}^TQ\mathbf{v})(\mathbf{v}^TQ^{-1}\mathbf{v})} \geqslant \frac{4 \cdot \lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}$$

With equality holding for some unit vector $\mathbf{v}$.
This can be rewritten:

$$\frac{(\mathbf{v}^T\mathbf{v})^2}{(\mathbf{v}^TQ\mathbf{v})} \geqslant (\mathbf{v}^TQ^{-1}\mathbf{v}) \cdot \frac{4 \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)}{\left(1 + \frac{\lambda_{\max}}{\lambda_{\min}}\right)^2}$$

We now proceed to the promised bound.

4.4 Main result: Bound on FLD in the projected space

Theorem 4.8. *Let $\mathbf{x_q} \sim \mathcal{D}_x = \mathcal{N}(\mu_y, \Sigma)$. Let $\mathcal{H}$ be the class of FLD functions and let $\hat{h}$ be the instance learned from the training data $\mathcal{T}_N$. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with entries drawn i.i.d from the univariate Gaussian $\mathcal{N}(0, 1/d)$. Then the estimated misclassification error $\hat{Pr}_{R,\mathbf{x}_q}[\hat{h}(R\mathbf{x}_q) \neq y]$ is bounded above by:*

$$\left(1 + \frac{1}{4} g(\hat{\Sigma}^{-1}\Sigma) \cdot \frac{1}{d} \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{\lambda_{\max}(\Sigma)}\right)^{-k/2} \qquad (4.20)$$

with μ_y the mean of the class from which $\mathbf{x}_q$ was drawn, estimated class means $\hat{\mu}_0$ and $\hat{\mu}_1$, model covariance $\hat{\Sigma}$, and $g(Q) = 4 \cdot \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \cdot \left(1 + \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}\right)^{-2}$.

Proof. We will start our proof in the dataspace, highlighting the contribution of covariance misspecification in the estimated error, and then make a move to the projected space with the use of a result (lemma 4.9) that shows that this component is always non-increasing under the random projection.

Without loss of generality we take $\mathbf{x}_q \sim \mathcal{N}(\mu_0, \Sigma)$, and for convenience take the estimated class priors to be equal i.e. $1 - \hat{\pi}_0 = \hat{\pi}_0$. By lemma 4.1, the estimated misclassification error in this case is upper bounded by:

$$\exp\left(-\frac{1}{8} \cdot \frac{\left[(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right) \qquad (4.21)$$

Now, in the Kantorovich inequality (lemma 4.7) we can take:

$$\mathbf{v} = \hat{\Sigma}^{-1/2}(\hat{\mu}_1 - \hat{\mu}_0)$$

where we use the fact ([13], Theorem 7.2.6, pg. 406) that since $\hat{\Sigma}^{-1}$ is symmetric positive definite it has a unique symmetric positive semi-definite square root:

$$\hat{\Sigma}^{-1/2} = \left(\hat{\Sigma}^{-1}\right)^{1/2} = \left(\hat{\Sigma}^{1/2}\right)^{-1} = \left(\hat{\Sigma}^{-1/2}\right)^T$$

and we will take our positive definite Q to be $Q = \hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}$ (ibid. pg. 406). Then, by lemma 4.7 we have the expression (4.21) is less than or equal to:

$$\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T\hat{\Sigma}^{-1/2}\left[\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}\right]^{-1}\hat{\Sigma}^{-1/2}(\hat{\mu}_1 - \hat{\mu}_0)\ldots\right.$$

$$\left.\ldots \times 4 \cdot \frac{\lambda_{\max}\left(\hat{\Sigma}^{-1}\Sigma\right)}{\lambda_{\min}\left(\hat{\Sigma}^{-1}\Sigma\right)} \cdot \left(1 + \frac{\lambda_{\max}\left(\hat{\Sigma}^{-1}\Sigma\right)}{\lambda_{\min}\left(\hat{\Sigma}^{-1}\Sigma\right)}\right)^{-2}\right) \qquad (4.22)$$

where the change in argument for the eigenvalues comes from the use of the identity *eigenvalues(AB) = eigenvalues(BA)* ([16], pg. 29). After simplification we can write this as:

$$\exp\left(-\frac{1}{8} \cdot (\hat{\mu}_1 - \hat{\mu}_0)^T\Sigma^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \cdot g(\hat{\Sigma}^{-1}\Sigma)\right) \qquad (4.23)$$

The term $g(\hat{\Sigma}^{-1}\Sigma)$ is a function of the model covariance misspecification, e.g. due to the imposition of diagonal or spherical constraints on $\hat{\Sigma}$. The following lemma shows that this term of the error can only decrease or stay the same after a random projection.

Lemma 4.9 (Non-increase of covariance misspecification error in the projected space)**.** *Let Q be a symmetric positive definite matrix. Let $K(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \in [1, \infty)$ be the reciprocal of the condition number of Q. Let $g(Q)$ be as given in the theorem 4.8. Then, for any fixed $k \times d$ matrix R with orthonormal rows:*

$$g((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T) \geqslant g(\hat{\Sigma}^{-1}\Sigma) \qquad (4.24)$$

Proof: We will show that $g(\cdot)$ is monotonic decreasing with K on $[1, \infty)$, then show that $K((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T) \leqslant K(\hat{\Sigma}^{-1}\Sigma)$, and hence $g((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T) \geqslant g(\hat{\Sigma}^{-1}\Sigma)$.

Step 1 *We show that g is monotonic decreasing:*
First note that for positive definite matrices $0 < \lambda_{\min} \leqslant \lambda_{\max}$, and so K is indeed in $[1, \infty)$. Differentiating $g(\cdot)$ with respect to K we get:

$$\frac{dg}{dK} = \frac{4(1 + K) - 8K}{(1 + K)^3} = \frac{4(1 - K)}{(1 + K)^3}$$

Here the denominator is always positive on the range of K while the numerator is always non-positive with *maximum 0 at $K = 1$. Hence $g(\cdot)$ is monotonic decreasing on $[1, \infty)$.*

Step 2 *We show that $K((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T) \leqslant K(\hat{\Sigma}^{-1}\Sigma)$: We will show that if $\hat{\Sigma}$ and Σ are symmetric positive definite and R is a random projection matrix then:*

$$\lambda_{\min}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{4.25}$$

$$\geqslant \lambda_{\min}(\hat{\Sigma}^{-1}\Sigma) = \lambda_{\min}(\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}) \tag{4.26}$$

and

$$\lambda_{\max}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{4.27}$$

$$\leqslant \lambda_{\max}(\hat{\Sigma}^{-1}\Sigma) = \lambda_{\max}(\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}) \tag{4.28}$$

Combining these inequalities then gives:

$$K((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T) \leqslant K(\hat{\Sigma}^{-1}\Sigma)$$

We give a proof of the first inequality, the second being proved similarly.

First, by lemma 4.4:

$$\lambda_{\min}([R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}) \tag{4.29}$$

$$= \min_{\mathbf{u}\in\mathbb{R}^k}\left\{\frac{\mathbf{u}^T[R\hat{\Sigma}R^T]^{-1/2}R\Sigma R^T[R\hat{\Sigma}R^T]^{-1/2}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}\right\} \tag{4.30}$$

Writing $\mathbf{v} = [R\hat{\Sigma}R^T]^{-1/2}\mathbf{u}$ so that $\mathbf{u} = [R\hat{\Sigma}R^T]^{1/2}\mathbf{v}$ then we may rewrite the expression (4.30), as the following:

$$= \min_{\mathbf{v}\in\mathbb{R}^k}\left\{\frac{\mathbf{v}^T R\Sigma R^T\mathbf{v}}{\mathbf{v}^T R\hat{\Sigma}R^T\mathbf{v}}\right\} \tag{4.31}$$

Writing $\mathbf{w} = R^T\mathbf{v}$, and noting that the span of all possible vectors $\mathbf{w}$ is a k-dimensional subspace of $\mathbb{R}^d$, we can bound the expression 4.31 by allowing the minimal vector $\mathbf{w} \in \mathbb{R}^d$ not to lie in this subspace:

$$\geqslant \min_{\mathbf{w}\in\mathbb{R}^d}\left\{\frac{\mathbf{w}^T\Sigma\mathbf{w}}{\mathbf{w}^T\hat{\Sigma}\mathbf{w}}\right\} \tag{4.32}$$

Now put $\mathbf{y} = \hat{\Sigma}^{1/2}\mathbf{w}$, with $\mathbf{y} \in \mathbb{R}^d$. This $\mathbf{y}$ exists uniquely since $\hat{\Sigma}^{1/2}$ is invertible, and we may rewrite (4.32) as the following:

$$= \min_{\mathbf{y}\in\mathbb{R}^d}\left\{\frac{\mathbf{y}^T\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right\} \tag{4.33}$$

$$= \lambda_{\min}(\hat{\Sigma}^{-1/2}\Sigma\hat{\Sigma}^{-1/2}) = \lambda_{\min}(\hat{\Sigma}^{-1}\Sigma) \tag{4.34}$$

This completes the proof of the first inequality, and a similar approach proves the second. Taken together the two inequalities imply $K(\hat{\Sigma}^{-1}\Sigma) \geqslant K([R\hat{\Sigma}R^T]^{-1}R\Sigma R^T)$ as required.

Finally putting the results of steps 1 and 2 together gives the lemma 4.9.

Back to the proof of theorem 4.8, we now move into the low dimensional space defined by any fixed random projection matrix R (i.e. with entries drawn from $\mathcal{N}(0, 1/d)$ and orthonormalised rows). By lemma 4.9, we can upper bound the projected space counterpart of (4.23) by the following:

$$\exp\left(-\frac{1}{8}\cdot(\hat{\mu}_1 - \hat{\mu}_0)^T R^T\left[R\Sigma R^T\right]^{-1}R(\hat{\mu}_1 - \hat{\mu}_0)\cdot g(\hat{\Sigma}^{-1}\Sigma)\right) \tag{4.35}$$

This holds for **any** fixed orthonormal matrix R, so it also holds for a fixed random projection matrix R.

Note, in the dataspace we bounded $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y]$ but in the projected space we want to bound:

$$\Pr_{R,\mathbf{x}_q}[\hat{h}(R\mathbf{x}_q) \neq y] = \mathrm{E}_{R,\mathbf{x}_q}[\mathcal{L}_{(0,1)}(\hat{h}(R\mathbf{x}_q), y)] \tag{4.36}$$

$$= \mathrm{E}_R[\mathrm{E}_{\mathbf{x}_q}[\mathcal{L}_{(0,1)}(\hat{h}(R\mathbf{x}_q), y)]|R] \tag{4.37}$$

This is the expectation of (4.35) w.r.t. the random choices of R. So we have:

$$\Pr_{R,\mathbf{x}_q}[\hat{h}(R\mathbf{x}_q) \neq y]$$

$$\leqslant \mathrm{E}_R\left[\exp\left(-\frac{1}{8}\cdot(\hat{\mu}_1 - \hat{\mu}_0)^T R^T\left[R\Sigma R^T\right]^{-1}R(\hat{\mu}_1 - \hat{\mu}_0)\ \ldots\right.\right.$$
$$\left.\left.\ldots \times g(\hat{\Sigma}^{-1}\Sigma)\right)\right] \tag{4.38}$$

$$\leqslant \mathrm{E}_R\left[\exp\left(-\frac{1}{8}\cdot\frac{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T R(\hat{\mu}_1 - \hat{\mu}_0)}{\lambda_{\max}(\Sigma)}\cdot g(\hat{\Sigma}^{-1}\Sigma)\right)\right] \tag{4.39}$$

where the last step is justified by lemma 4.6. Now, since the entries of R where drawn i.i.d from $\mathcal{N}(0, 1/d)$, the term $(\hat{\mu}_1 - \hat{\mu}_0)^T R^T R(\hat{\mu}_1 - \hat{\mu}_0) = \|R(\hat{\mu}_1 - \hat{\mu}_0)\|^2$ is χ_k^2 distributed and (4.39) is therefore the moment generating function of a χ^2 distribution.

Hence we can rewrite (4.39) as:

$$= \left[1 + \frac{1}{4}\cdot g(\hat{\Sigma}^{-1}\Sigma)\cdot\frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d\cdot\lambda_{\max}(\Sigma)}\right]^{-k/2} \tag{4.40}$$

A similar sequence of steps proves the other side, when $\mathbf{x}_q \sim \mathcal{N}(\mu_1, \Sigma)$, and gives the same expression. Then putting the two terms together, applying the law of total probability with $\sum_{y\in\mathcal{C}}\pi_y = 1$ finally gives theorem 4.8.

4.4.1 Comment: Other projections R

Although we have taken the entries of R be drawn from $\mathcal{N}(0, 1/d)$ this was used only in the final step, in the form of the moment generating function of the χ^2 distribution. In consequence, other distributions that produce inequality in the step from equation (4.39) to equation (4.40) suffice. Such distributions include sub-Gaussians and some examples of suitable distributions may be found in [1]. Whether any deterministic projection R can be found that is both non-adaptive (i.e. makes no use of the training labels) and still yields a non-trivial guarantee for FLD in terms of only the data statistics seems a difficult open problem.

4.5 Bound on the projected dimensionality k and discussion

For both practical and theoretical reasons, we would like to know to which dimensionality k we can project our original high dimensional data and still expect to recover good classification performance from FLD. This may be thought of as a measure of the difficulty of the classification task.

By setting our bound to be no more than $\delta \in (0, 1)$ and solving for k we can obtain such a bound on k for FLD that guarantees that the expected misclassification probability (w.r.t. R) in the projected space remains below δ:

Corollary 4.10. *[to Theorem 4.8] Let k, d, $g(\cdot)$, $\hat{\mu}_y$, Σ, $\hat{\Sigma}$, $\mathcal{C}$ be as given in theorem 4.8. Then, in order that the probability of misclassification in the projected space remains below δ it is sufficient to take:*

$$k \geqslant 8 \cdot \frac{d\lambda_{\max}(\Sigma)}{\min\limits_{i,j\in\mathcal{C},i\neq j}\|\hat{\mu}_i - \hat{\mu}_j\|^2} \cdot \frac{1}{g(\hat{\Sigma}^{-1}\Sigma)} \cdot \log(m/\delta) \qquad (4.41)$$

Proof: In the 2-class case we have:

$$\delta \geqslant \left[1 + \frac{1}{4} \cdot g(\hat{\Sigma}^{-1}\Sigma) \cdot \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d\lambda_{\max}(\Sigma)}\right]^{-k/2} \iff \qquad (4.42)$$

$$\log(1/\delta) \leqslant \frac{k}{2}\log\left[1 + \frac{1}{4} \cdot g(\hat{\Sigma}^{-1}\Sigma) \cdot \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d\lambda_{\max}(\Sigma)}\right] \qquad (4.43)$$

then using the inequality $(1 + x) \leqslant e^x$, $\forall\ x \in \mathbb{R}$ we obtain:

$$k \geqslant 8 \cdot \frac{d\lambda_{\max}(\Sigma)}{\|\hat{\mu}_1 - \hat{\mu}_0\|^2} \cdot \frac{1}{g(\hat{\Sigma}^{-1}\Sigma)} \cdot \log(1/\delta) \qquad (4.44)$$

Using (4.44) and lemma 4.2, it is then easy to see that to expect no more than δ error from FLD in an $m + 1$-class problem, the required dimension of the projected space need only be:

$$k \geqslant 8 \cdot \frac{d\lambda_{\max}(\Sigma)}{\min_{i,j\in\mathcal{C},i\neq j}\|\hat{\mu}_i - \hat{\mu}_j\|^2} \cdot \frac{1}{g(\hat{\Sigma}^{-1}\Sigma)} \cdot \log(m/\delta) \quad (4.45)$$

as required.

We find it interesting to compare our k bound with that given in the seminal paper of Arriaga and Vempala [3]. The analysis in [18] shows that the bound in [3] for randomly projected 2-class perceptron classifiers is equivalent to requiring that the projected dimensionality

$$k = \mathcal{O}\left(72 \cdot \frac{L}{l^2} \cdot \log(6N/\delta)\right) \qquad (4.46)$$

where δ is the user-specified tolerance of misclassification probability, N is the number of training examples, and L/l^2 is the diameter of the data ($L = \max_{n=1,\ldots,N}\|\mathbf{x}_n\|^2$) divided by the margin (or 'robustness', as they term it). In our bound, $g(\cdot)$ is a function that encodes the quality of the model covariance specification, δ and k are the same as in [3] and the factor $d\lambda_{\max}(\Sigma) \cdot \|\hat{\mu}_1 - \hat{\mu}_0\|^{-2}$ — which, should be noted, is exactly the reciprocal of the squared class separation as defined by Dasgupta in [6] — may be thought of as the 'generative' analogue of the data diameter divided by the margin in (4.46).

Observe, however, that (4.46) grows with the log of the training set size, whereas ours (4.44) grows with the log of the number of classes. This is not to say, by any means, that FLD is superior to perceptrons in the projected space. Instead, the root and significance of this difference lies in the assumptions (and hence the methodology) used in obtaining the bounds. The result in (4.46) was derived from the precondition that all pairwise distances between the training points must be approximately preserved *uniformly* cf. the Johnson-Lindenstrauss lemma [8]. It is well understood [2] that examples of data sets exist for which the $k = \mathcal{O}(\log N)$ dimensions are indeed required for this. However, we conjecture that, for learning, this starting point is too strong a requirement. Learning should not become harder with more training points — assuming of course that additional examples add 'information' to the training set.

Our derivation is so far specific to FLD, but it is able to take advantage of the class structure inherent in the classification setting in that the misclassification error probability is down to very few key distances only — the ones between the class centers.

Despite this difference from [3] and approaches based on uniform distance preservation, in fact our conclusion should not be too surprising. Earlier work in theoretical computer science [6] proves performance guarantees with high probability (over the choice of R) for the *unsupervised* learning of a mixture of Gaussians which also requires k to grow logarithmically with the number of classes only. Moreover, our finding that the error from covariance misspecification is always non-increasing in the projection space is also somewhat expected, in the light of the finding in [6] that projected covariances tend to become more spherical.

In closing, it is also worth noting that the extensive empirical results in e.g. [7] and [11] also suggest that classification (including non-sparse data) requires a much lower projection dimension than that which is needed for global preservation of all pairwise distances cf. the JLL. We therefore conjecture that, all other things being equal, the difficulty of a classification task should be a function only of selected distances, and preserving those may be easier that preserving every pairwise distance uniformly. Investigating this more generally remains for further research.

4.6 Numerical validation

We present three numerical tests that illustrate and confirm our main results.

Lemma 4.9 showed that the error contribution of a covariance misspecification is always no worse in the low dimensional space than in the high dimensional space. Figure 1 shows the quality of fit between a full covariance Σ and its diagonal approximation $\hat{\Sigma}$ when projected from a $d = 100$ dimensional data space into successively lower dimensions k. We see the fit is poor in the high dimensional space, and it keeps improving as k gets smaller. The error bars span the minimum and maximum of $g([R\hat{\Sigma}R^T]^{-1}R\Sigma R^T)$ observed over 40 repeated trials for each k.

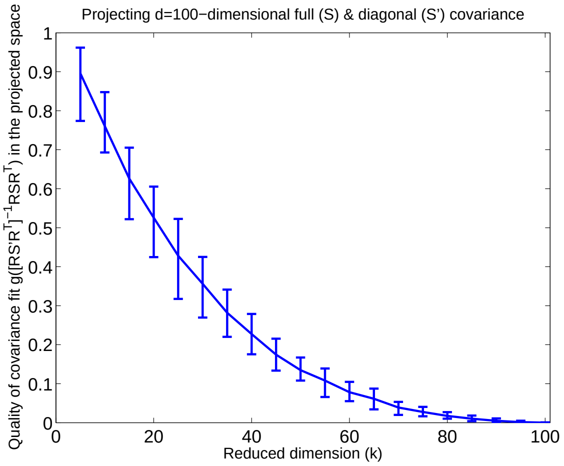

Figure 1: Experiment confirming Lemma 4.9; the error contribution of a covariance misspecification is always no worse in the projected space than in the data space. The best possible value on the vertical axis is 1, the worst is 0. We see the quality of fit is poor in high dimensions and improves dramatically in the projected space, approaching the best value as k decreases.

The second set of experiments demonstrates Corollary 4.10 of our Theorem 4.8, namely that for good generalisation of FLD in the projected space, the required projection dimension k is logarithmic in the number of classes.

We randomly projected m equally distanced spherical unit variance 7-separated Gaussian classes from $d = 100$ dimensions and chose the target dimension of the projected space as $k = 12\log(m)$. The boxplots in figure 2 show, for each m tested, the distribution of the empirical error rates over 100 random realisations of R, where for each R the empirical error was estimated from 500 independent query points. Other parameters being unchanged, we see the classification performance is indeed maintained with this choice of k.

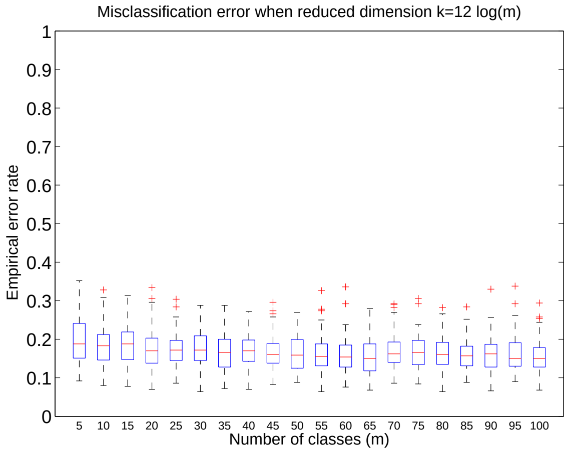

Figure 2: Experiment illustrating Theorem 4.8 & its Corollary 4.10. With the choice $k = 12\log(m)$ and $\|\mu_i - \mu_j\| = 7$, $\forall i \neq j$, the classification performance is kept at similar rates while the number of classes m varies.

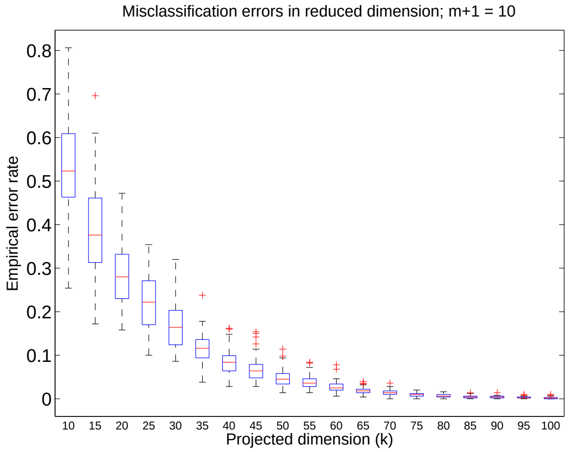

Figure 3: Experiment illustrating Theorem 4.8. We fix the number of classes, $m + 1 = 10$, and the data dimensionality, $d = 100$, and vary the projection dimensionality k. The classification error decreases nearly exponentially as $k \to d$.

The third experiment shows the effect of reducing k for a 10-class problem in the same setting as experiment two. As expected, the classification error in figure 3 decreases nearly exponentially as the projected dimensionality k tends to the data dimensionality d. We note also, from these empirical results, that the variability in the classification performance also decreases with increasing k. Finally, we observe that the worst performance in the worst case is still a weak learner that performs better than chance.

4.7 Two straightforward extensions

We proved our theorem 4.8 for classes with identical covariance matrices in order to ensure that our exposition was reasonably sequential, as well as to keep our notation as uncluttered as possible. However, it can be seen from the proof that this is not essential to our argument and, in particular, the following two simple extensions can be easily proved:

4.7.1 Different unimodal classes

By replacing Σ in equation (7.1) with Σ_y, the following analogue of theorem 4.8 can be derived for the 2-class case when the true class structure is Gaussian (or sub-Gaussian) but with different class covariance matrices Σ_0 and Σ_1:

$$\pi_0 \left[1 + \frac{1}{4} \cdot g(\hat{\Sigma}^{-1}\Sigma_0) \cdot \frac{\|\hat{\mu}_1 - \hat{\mu}_0\|^2}{d \cdot \lambda_{\max}(\Sigma_0)}\right]^{-k/2} + \dots$$

$$\dots \pi_1 \left[1 + \frac{1}{4} \cdot g(\hat{\Sigma}^{-1}\Sigma_1) \cdot \frac{\|\hat{\mu}_0 - \hat{\mu}_1\|^2}{d \cdot \lambda_{\max}(\Sigma_1)}\right]^{-k/2}$$

4.7.2 Different multimodal classes

In a similar way if the classes have a multimodal structure then, by representing each class as a finite mixture of Gaussians $\sum_{i=1}^{M_y} w_{yi}\mathcal{N}(\mu_{yi}, \Sigma_{yi})$, it is not too hard to see that provided the conditions of theorem 4.8 hold for the Gaussian in the mixture with the *greatest* contribution to the misclassification error, we can upper bound the corresponding form of equation (7.1), for the case $y = 0$ as follows:

$$\mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\mathbf{x}_q\right)\right] =$$

$$\sum_{i=1}^{M_0} w_{0i} \exp\bigg(\mu_{0i}^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \dots$$

$$\dots + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1}\Sigma_{0i}\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\bigg)$$

$$\leqslant \max_{i \in \{1,\dots,M_0\}} \bigg\{\exp\bigg(\mu_{0i}^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \dots$$

$$\dots + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1}\Sigma_{0i}\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\bigg)\bigg\}$$

Where M_y is the number of Gaussians in the mixture for class y, w_{yi} is the weight of the i-th Gaussian in the mixture, $\sum_i w_{yi} = 1$, and μ_{yi}, Σ_{yi} are the corresponding true mean and true covariance.

The proof then proceeds as before and the resultant bound, which of course is nowhere near tight, still gives k of the same order as the bound in theorem 4.8. In this setting, the condition $\kappa_y > 0$ implies that the centres of the Gaussian mixture components are at least nearly linearly separable. In the high-dimensional data space this can still be a reasonable assumption (unless the number of mixture components is large), but it is clear that in practice it is much less likely to hold in the projected space.

5. CONCLUSIONS

We considered the problem of classification in non-adaptive dimensionality reduction using FLD in randomly projected data spaces.

Previous results considering other classifiers in this setting gave guarantees on classification performance only when all pairwise distances were approximately preserved under projection. We conjectured that, if one were only interested in preserving classification performance, that it would be sufficient to preserve only certain key distances. We showed that, in the case of FLD, this is sufficient (namely preserving the separation of the class means). We employed a simple generative classifier in our working, but one might imagine that e.g. for projected SVM it would be sufficient to preserve only the separation of the support vectors. Our only assumption on the data was that the distribution of data points in each class be dominated by a Gaussian and, importantly, we did not require our data to have a sparse or implicity low-dimensional structure in order to preserve the classification performance. We also showed that misspecification of the covariance of the Gaussian in the projected space has a relatively benign effect, when compared to a similar misspecification in the original high dimensional data space, and we proved that if $k = \mathcal{O}\log(m)$ then it is possible to give guarantees on the expected classification performance (w.r.t R) of projected FLD.

One practical consequence of our results, and the other similar results in the literature, is to open the door to the possibility of collecting and storing data in low-dimensional form whilst still retaining guarantees on classification performance.

Moreover, answering these questions means that we are able to foresee and predict the behaviour of a randomly projected FLD classifier, and the various factors that govern it, before actually applying it to a particular data set.

Future research includes an analysis of the behaviour of the projected classifier when there is only a small amount of training data, and the extension of our results to general multimodal classes.

We note that the finite sample effects are characterised by (i) the estimation error and (ii) the fact that when the condition $\kappa_y > 0$ holds in the data space it is possible that random projection causes it to no longer hold in the projected space.

The estimation error (i) depends on the quality of the estimates of the class means and class covariances, and these can be analysed using techniques that are not specific to working in the randomly projected domain, e.g. [14, 17]. Observe, however, that because there are fewer parameters to estimate, the estimation error in the projected space must be less than the estimation error in the data space.

The second effect (ii) involves the probability that two vectors in the data space with angular separation $\theta \in (0, \pi/2)$, have angular separation $\theta_R > \pi/2$ in the projected space following random projection. We can show that this probability is typically small, and our results regarding this effect are currently in preparation for publication [10].

6. REFERENCES

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[2] N. Alon. Problems and results in extremal combinatorics, Part I. *Discrete Math*, 273:31–53, 2003.

[3] R. Arriaga and S. Vempala. An algorithmic theory of learning. *Machine Learning*, 63(2):161–182, 2006.

[4] P. Bickel and E. Levina. Some theory for Fisher's linear discriminant function, 'naïve Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

[5] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University, 2009.

[6] S. Dasgupta. Learning Mixtures of Gaussians. In *Annual Symposium on Foundations of Computer Science*, volume 40, pages 634–644, 1999.

[7] S. Dasgupta. Experiments with random projection. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 143–151, 2000.

[8] S. Dasgupta and A. Gupta. An Elementary Proof of the Johnson-Lindenstrauss Lemma. *Random Struct. Alg.*, 22:60–65, 2002.

[9] M.A. Davenport, M.B. Wakin and R.G. Baraniuk. Detection and estimation with compressive measurements. Technical Report TREE 0610, Rice University, January 2007.

[10] R.J. Durrant and A. Kabán. Finite Sample Effects in Compressed Fisher's LDA. Unpublished 'breaking news' poster presented at AIStats 2010, www.cs.bham.ac.uk/∼durranrj.

[11] D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 522. ACM, 2003.

[12] J. Haupt, R. Castro, R. Nowak, G. Fudge and A. Yeh. Compressive sampling for signal classification. In *Proc. 40th Asilomar Conf. on Signals, Systems, and Computers*, pages 1430–1434, 2006.

[13] R.A. Horn and C.R. Johnson. *Matrix Analysis*. CUP, 1985.

[14] O.-A. Maillard and R. Munos. Compressed Least-Squares Regression. In *NIPS*, 2009.

[15] T. Pattison and D. Gossink. Misclassification Probability Bounds for Multivariate Gaussian Classes. *Digital Signal Processing*, 9:280–296, 1999.

[16] K.B. Petersen and M.S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, November 2008.

[17] J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

[18] T. Watanabe, E. Takimoto, K. Amano and A. Maruoka. Random projection and its application to learning. In *Randomness and Computation Joint Workshop 'New Horizons in Computing' and 'Statistical Mechanical Approach to Probabilistic Information Processing'*, July 2005.

7. APPENDIX

PROOF. (of lemma 4.1) We prove one term of the bound using standard techniques, the other term being proved similarly.
Without loss of generality let $\mathbf{x}_q$ have label $y = 0$. Then the probability that $\mathbf{x}_q$ is misclassified is given by $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y | y = 0] = \Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0]$:

$$=\Pr_{\mathbf{x}_q}\left[\log\frac{1-\hat{\pi}_0}{\hat{\pi}_0} + (\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right]$$
$$=\Pr_{\mathbf{x}_q}\left[\kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0} + (\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right]$$

for all $\kappa_0 > 0$. Exponentiating both sides gives:

$$=\Pr_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) + \kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right) > 1\right]$$
$$\leqslant \mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}\left(\mathbf{x}_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) + \kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right)\right]$$

by Markov inequality. Then, isolating terms in $\mathbf{x}_q$ we have $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0] \leqslant$

$$\mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1} \mathbf{x}_q \ldots\right.\right.$$
$$\left.\left.\ldots - \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1) + \kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right)\right]$$
$$= \exp\left(-\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1) + \kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right)\ldots$$
$$\ldots \times \mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1} \mathbf{x}_q\right)\right]$$

This expectation is of the form of the moment generating function of a multivariate Gaussian and so:

$$\mathrm{E}_{\mathbf{x}_q}\left[\exp\left((\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1} \mathbf{x}_q\right)\right] =$$
$$\exp\left(\mu_0^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\right) \tag{7.1}$$

where μ_0 is the true mean, and Σ is the true covariance matrix, of $\mathcal{D}_{\mathbf{x}_q}$.

Thus, we have the probability of misclassification is bounded above by the following:

$$\exp\left(-\frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_0 + \hat{\mu}_1) + \kappa_0 \log\frac{1-\hat{\pi}_0}{\hat{\pi}_0} \ldots\right.$$
$$\left.\ldots + \mu_0^T \kappa_0 \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_0)^T \kappa_0^2 \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)\right)$$

Now, since this holds for every $\kappa_0 > 0$ we may optimise the bound by choosing the best one. Since exponentiation is a monotonic increasing function, in order to minimise the bound it is sufficient to minimise its argument. Differentiating the argument w.r.t κ_0 and setting the derivative equal to zero then yields:

$$\kappa_0 = \frac{(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - 2\log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}}{2(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)} \tag{7.2}$$

This is strictly positive as required, since the denominator is always positive (Σ is positive definite, then so is $\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}$), and the numerator is assumed to be positive as a precondition in the theorem.

Substituting κ_0 back into the bound then yields, after some algebra, the following $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 0] \leqslant$

$$\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - 2\log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right)$$

The second term, for when $\mathbf{x}_q \sim \mathcal{N}(\mu_1, \Sigma)$, can be derived similarly and gives $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq 1] \leqslant$

$$\exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_1)^T \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1) - 2\log\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right]^2}{(\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}\right)$$

Finally, putting these two terms together and applying the law of total probability (since the classes in $\mathcal{C}$ partition the data): $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] = \sum_{y \in \mathcal{C}} \Pr[\mathbf{x}_q \sim \mathcal{N}(\mu_y, \Sigma)] \cdot \Pr[\hat{h}(\mathbf{x}_q) \neq y | \mathbf{x}_q \sim \mathcal{N}(\mu_y, \Sigma)]$, we arrive at lemma 4.1, i.e. that $\Pr_{\mathbf{x}_q}[\hat{h}(\mathbf{x}_q) \neq y] \leqslant$

$$\pi_0 \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - 2\log\frac{1-\hat{\pi}_0}{\hat{\pi}_0}\right]^2}{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}\right) +$$
$$(1-\pi_0) \exp\left(-\frac{1}{8}\frac{\left[(\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_1)^T \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1) - 2\log\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right]^2}{(\hat{\mu}_0 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}\right)$$

7.0.3 *Comment 1*

We should confirm, of course, that the requirement that $\kappa_y > 0$ is a reasonable one. Because the denominator in (7.2) is always positive the condition $\kappa_y > 0$ holds when:

$$(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)^T \hat{\Sigma}^{-1}(\hat{\mu}_{\neg y} - \hat{\mu}_y) - 2\log\left(\frac{1-\hat{\pi}_y}{\hat{\pi}_y}\right) > 0$$

It can be seen that $\kappa_y > 0$ holds provided that for each class the true and estimated means are both on the same side of the decision hyperplane. Furthermore, provided that $\kappa_y > 0$ holds in the data space we can show that w.h.p (and independently of the original data dimensionality d) it also holds in the projected space [10], and so the requirement $\kappa_y > 0$ does not seem particularly restrictive in this setting.

7.0.4 *Comment 2*

We note that, in (7.1) it is in fact sufficient to have inequality. Therefore our bound also holds when the true distributions $\mathcal{D}_x$ of the data classes are such that they have a moment generating function no greater than that of the Gaussian. This includes sub-Gaussian distributions, i.e. distributions whose tail decays faster than that of the Gaussian.