UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

The effect of intrinsic dimension on the Bayes-error of projected quadratic discriminant classification

Palias, Efstratios; Kabán, Ata

DOI:
10.1007/s11222-023-10251-1

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Palias, E & Kabán, A 2023, 'The effect of intrinsic dimension on the Bayes-error of projected quadratic discriminant classification', *Statistics and Computing*, vol. 33, no. 4, 87. https://doi.org/10.1007/s11222-023-10251-1

Link to publication on Research at Birmingham portal

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 03. Aug. 2023

Statistics and Computing (2023) 33:87
https://doi.org/10.1007/s11222-023-10251-1

ORIGINAL PAPER

The effect of intrinsic dimension on the Bayes-error of projected quadratic discriminant classification

Efstratios Palias[1] · **Ata Kabán**[1]

Received: 31 January 2023 / Accepted: 28 April 2023
© The Author(s) 2023

Abstract
High-dimensionality is a common hurdle in machine learning and pattern classification; mitigating its effects has attracted extensive research efforts. It has been found in a recent NeurIPS paper that, when the data possesses a low effective dimension, the predictive performance of a discriminative quadratic classifier with nuclear norm regularisation enjoys a reduced (logarithmic) dependence on the ambient dimension and depends on the effective dimension instead, while other regularisers are insensitive to the effective dimension. In this paper, we show that dependence on the effective dimension is also exhibited by the Bayes error of the generative Quadratic Discriminant Analysis (QDA) classifier, without any explicit regularisation, under three linear dimensionality reduction schemes. Specifically, we derive upper bounds on the Bayes error of QDA, which adapt to the effective dimension, and entirely bypass any dependence on the ambient dimension. Our findings complement previous results on compressive QDA that were obtained under compressive sensing type assumptions on the covariance structure. In contrast, our bounds make no a-priori assumptions on the covariance structure, in turn they tighten in the presence of benign traits of the covariance. We corroborate our findings with numerical experiments.

Keywords Dimensionality reduction · Quadratic discriminant analysis · High-dimensional classification · Bhattacharyya bound

1 Introduction

Modern data sets often consist of large numbers of features. When the data is high dimensional, the "curse of dimensionality" typically degrades the performance of machine learning algorithms, as learning in high dimensions requires much larger training sample sizes and increased computational resources [1]. It has been observed however that many real-life data sets do not fill their ambient spaces evenly [2, 3], in which case, learning form them is expected to be easier both statistically and computationally. To distinguish this structural notion of inherent dimension of data from the ambient dimension, we shall refer to it, in a general sense, as the intrinsic dimension (ID). We thus refer to such data sources as possessing a low-ID.

A large volume of research has been dedicated to mitigating the ill effects of high dimensionality, including regularisation methods, and dimensionality reduction approaches. There is also an increasing number of studies aimed at elucidating whether some definition of ID allows learning algorithms to succeed with less resources [3–7], mostly focusing on complex models.

At the confluence of regularisation-based approaches and intrinsic dimension-based analyses, a recent study at NeurIPS [8] examined homogeneous discriminative quadratic classifiers, and proved that, whenever the data has a low effective dimension (a notion of ID), a nuclear norm constraint of its matrix parameter enables it to adapt to the effective dimension of the data, and have only logaritmic dependence on the ambient dimension. In other words, the classifier obtained needs less training data to achieve good generalisation whenever the data has a low effective dimension. The authors also confirmed this effect experimentally, and demonstrated that a number of other norm-constraint based regularisation schemes lack such desirable property.

Motivated by these recent results, our goal in this paper is to investigate the effect of intrinsic or effective dimension in

✉ Efstratios Palias
exp093@bham.ac.uk

Ata Kabán
a.kaban@bham.ac.uk

[1] School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Published online: 02 June 2023

QDA, without explicit regularisation, under dimensionality reduction schemes.

Dimensionality reduction essentially means to extract or to construct a smaller number of features than the number of observed or measured ones. Its aim is to improve computational efficiency, while potentially sacrificing the accuracy a little, or even enhancing accuracy by reducing noise. Some methods examine the data set to find its key features, others perform a general purpose compression. Linear projections are popular approaches, whereby the original high dimensional data is linearly projected into a lower dimensional subspace before pursuing a learning task. Examples include Random Projections (RP), and Principal Components Analysis (PCA).

Our interest in dimension-reducing projections stems from the fact that, for some algorithms, they can reduce or eliminate the ambient dimension dependence of their accuracy. For instance, for the popular classification algorithm of Linear Discriminant Analysis (LDA), it has been found that, under mild assumptions, the generalisation error of its randomly projected version does not depend on the ambient dimension [9, 10]. In addition, a significant advantage of RP is that it can take advantage of low-ID without the need to know or compute the ID of the data.

LDA is special in that its Bayes-error has a closed form. It is natural to ask if a similar behaviour can be found in the more general classification algorithm of Quadratic Discriminant Analysis (QDA). This is what we set out to study in this paper. The importance of our work stems from the popularity of QDA as a classification algorithm, with many applications throughout several decades, including genetics [11, 12], medicine [13, 14] and image processing [15, 16].

The remainder of this paper is structured as follows: In Sect. 2, we formally introduce the model of QDA. We then review the related literature and state our contributions. In Sect. 3, we present and discuss our results and findings. Section 4 is devoted to series of controlled experiments on synthetic data, designed to verify whether the bounds we derived reveal something real about the effect of a low-ID on the generalisation error of QDA. Section 5 concludes the paper and outlines possible future research directions.

2 Background

Linear and Quadratic Discriminant Analysis (LDA and QDA) were introduced by Fisher in 1936 [17] as classification algorithms for two or more classes of instances. These still represent popular methods in statistical pattern recognition due to their effectiveness, simplicity and interpretability [18, 19]. Due to their success, they have been extensively studied over the past decades. However, the bulk of analytic work is focused on LDA, as its Bayes-error has a closed-form expression. No closed form is available for the Bayes error of QDA, its analysis is harder, and existing results are much more scarce. In this work we focus on QDA.

2.1 The QDA model

Quadratic Discriminant Analysis (QDA) is a generative classifier. Let $X \times Y$ be the instance space, where X is the feature space and $Y = \{0, 1\}$ is the set of labels. We will write each instance as a pair (x, y), where $x \in X$ is the feature vector and $y \in Y$ is its label. Let $\mathcal{C}_0$ and $\mathcal{C}_1$ denote the two data classes. QDA assumes class-conditional Gaussian models on both classes. That is, the data instances are assumed to be generated i.i.d. as follows.

1. First, one of the two classes is chosen randomly with probabilities π_0 for $\mathcal{C}_0$ and π_1 for $\mathcal{C}_1$, where π_0 and π_1 are class priors, and satisfy $0 < \pi_0, \pi_1 < 1$ and $\pi_0 + \pi_1 = 1$;
2. For the chosen class, the datum instance is generated from the class-specific Gaussian distribution, which is $\mathcal{N}(\mu_0, \Sigma_0)$ for $\mathcal{C}_0$ and $\mathcal{N}(\mu_1, \Sigma_1)$ for $\mathcal{C}_1$.

Here, $\mathcal{N}(\mu, \Sigma)$ denotes the multi-variate Gaussian distribution with mean vector μ and covariance matrix Σ. LDA assumes further that $\Sigma_0 = \Sigma_1$, while QDA does not make this assumption. Thus, the parameters of QDA are the following:

$$\pi_0, \mu_0, \Sigma_0, \pi_1, \mu_1, \Sigma_1. \tag{1}$$

In practice, the true parameters are not available, so we usually compute estimates from a finite sample or training set. However, this paper is concerned with the Bayes-error of QDA, which is a theoretical quantity representing the expected misclassification error as a function of the true parameters. We will therefore assume that the true parameters in (1) are known.

For a given feature vector $x \in X$, and the true parameter values, the prediction rule of QDA is a function $f : X \to Y$ defined as

$$f(x) = \begin{cases} 0 \text{ if } \pi_0 \cdot p(x|\mu_0, \Sigma_0) \geq \pi_1 \cdot p(x|\mu_1, \Sigma_1); \\ 1 \text{ otherwise,} \end{cases} \tag{2}$$

where $p(x|\mu, \Sigma)$ is the density of $\mathcal{N}(\mu, \Sigma)$ evaluated at the point x. Expanding out the definition of the Gaussian density it is easy to see that this prediction rule takes the form of a quadratic decision boundary in general. The model reduces to LDA when $\Sigma_0 = \Sigma_1$, in which case the decision boundary is linear [20].

The generalisation error of any classifier is the probability of misclassification. For QDA, this is

$$\Pr_{(x,y)\sim\mathcal{D}}\{f(x) \neq y\} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbb{I}\{f(x) \neq y\}], \tag{3}$$

where $\mathcal{D} := \pi_0 \cdot \mathcal{N}(\mu_0, \Sigma_0) + \pi_1 \cdot \mathcal{N}(\mu_1, \Sigma_1)$ is the data generating distribution under the QDA model, $\mathbb{I}\{\cdot\}$ is the indicator function, which equals 1 if its argument is true, and 0 otherwise, and f is defined as in (2). This is the probability that a future instance from either class is incorrectly classified. The ultimate goal of any machine learning algorithm to minimize this quantity.

Under the QDA model, there is a well-known upper bound on the Bayes error of QDA, the Bhattacharyya bound, which will be our starting point in this paper. To simplify the setting, we will consider the common case where the class priors π_0 and π_1 are equal.

Theorem 1 (Bhattacharyya bound [18]) *Let* $\mu_0, \mu_1 \in \mathbb{R}^d$, $\Sigma_0, \Sigma_1 \in \mathbb{S}^d_+$. *Then, an upper bound for the generalisation error of* (3) *is given by* $\exp(-\delta(\mu_0, \Sigma_0, \mu_1, \Sigma_1))$, *where*

$$\delta(\mu_0, \Sigma_0, \mu_1, \Sigma_1) := \frac{1}{8}(\mu_0 - \mu_1)^T \left(\frac{\Sigma_0 + \Sigma_1}{2}\right)^{-1} (\mu_0 - \mu_1) + \frac{1}{2}\log \frac{\left|\frac{\Sigma_0 + \Sigma_1}{2}\right|}{\sqrt{|\Sigma_0|\,|\Sigma_1|}}. \tag{4}$$

The expression (4) is called the Bhattacharyya distance (hence the name of the bound), as it is a measure of statistical distance between two Gaussian distributions, being zero if and only if the two distributions have the same parameters [21].

2.2 Related work

Our work is focused on the generalisation error of QDA under projection based dimensionality reduction. We want to find out whether it adapts to some notion of low-ID in the data. This question was previously studied in the simpler setting of LDA [9, 10], where the authors give high-probability upper bounds on the error of LDA under Gaussian random projection. In the finite dimensional case, if the true covariance matrix has a low rank, they find that the ambient dimension is replaced with this rank. Moreover, in the infinite-dimensional case, if the covariance has a finite trace, the dimension is replaced with another notion of ID known as the effective dimension. This is the ratio of the trace of the covariance matrix and its largest eigenvalue. Interestingly, the same notion of ID appears in the analysis of nuclear-norm regularised discriminative quadratic classifier in the recent study of [8]. This particular notion of ID will also appear in our results, so we will define it formally in Sect. 4.

As we already mentioned, similar results are sparse for QDA. The recent study at Neurips [8] considered a specific form of quadratic classifiers, namely homogeneous discriminative quadratic classifiers of the form $x^T Ax \gtrless 0$, where A can be an indefinite matrix parameter, and is subjected to a nuclear norm regularisation constraint. Their results improve on the then-best known bounds from [22] by reducing the dependence on the ambient dimension to logarithmic, and showing the ability of the nuclear norm regulariser to adapt to the intrinsic dimension. They also demonstrate experimentally that several other regularisers, such as the Frobenius norm regulariser, lack this ability.

The main analytic tool used in [8] is the Rademacher complexity, which has the advantage of distribution-freeness (no class-conditional Gaussian assumption). However, they do not consider the full generative QDA model, and do not consider any dimensionality reduction method, but only the discriminative classifier that operates in the full ambient space. Moreover, their bound still has a logarithmic dependence on the ambient dimension. It is therefore of interest to us to consider the full generative QDA model, and consider dimensionality reduction rather than regularisation. In QDA, the matrix that appears in the decision boundary can be positive- or negative-definite, and the quadratic expression in the decision boundary is not necessarily homogeneous.

Another thread of work similar to ours appeared in [23], where a multi-class Gaussian classifier very similar to QDA is analysed under randomly projected measurement data. However, the authors assume that the original high dimensional covariances are low-rank, and have some Gaussian noise added after being randomly projected. Furthermore, the covariance matrices are assumed to be known and span random subspaces of equal dimensions. These subspaces are assumed to have been sampled randomly uniformly, in order to guarantee that the average of two covariance matrices has double the rank of the individual matrices almost surely. These assumptions originate from the field of compressed sensing, and help obtaining results though techniques inspired from that field.

Among other results, they find that, under their assumptions, the minimum number of measurements to guarantee a reliable classification performance depends on the rank of the covariance matrices instead of the ambient dimension. This is another step towards quantifying the dependence of classification on the ID. However, the restrictions on the covariance matrices being low-rank and also of equal ranks are rather unrealistic, and most of their results degrade rapidly when the matrices are of "almost" low-rank. The equal- and zero-means case, in the ambient space is also quite restrictive for real-world data.

The imposition of such strong assumptions means that the results only hold when the assumptions hold. Instead, we aim to forego such a-priori structural assumptions where possible, and let our bounds reveal benign structural traits where the bounds tighten. Interestingly, we shall see good agreement between the benign structures identified this way from our bounds and those assumed a-priori in [23]; however our bounds are still valid without those a-priori assumptions.

We consider three dimensionality-reducing projections owing to their popularity, simplicity, and their diversity on how they operate on the data. The first two are random Gaussian and random orthogonal projections; these both are oblivious to the parameters and data, while the third one is principal component analysis (PCA), a deterministic function of the data parameters. To the best of our knowledge, there are no results on the error of QDA under projections that would completely eliminate dependence on the ambient dimension without a strong low-rank assumption. We are also unaware of work quantifying the effect of PCA on the error of QDA.

Furthermore, it is well known that QDA in a large ambient space suffers from the curse of dimensionality. It is therefore imperative to study its classification performance under several different dimensionality reduction schemes, and unveil its degree of dependence from the ambient dimension, at least theoretically. This would shed light and improve our understanding of how we can expect the algorithm to perform when faced with high-dimensional data sets, especially when they possess a low-ID. Motivated by this goal, we revisit QDA under projections, without imposing any assumptions other than the QDA model itself.

2.3 Our contributions

We summarise our contributions as follows.

1. We upper bound the Bayes error of QDA under three popular projection schemes: Gaussian random projection; random orthogonal projection; and principal components analysis.
   - Our bounds eliminate the dependence on the ambient dimension and replace it with a specific notion of ID, i.e. the effective dimension.
   - Our bounds highlight some benign structural traits of the problem, including favourable covariance structures, without imposing any such structures a-priori.
2. We corroborate our theoretical findings with controlled numerical experiments on synthetic data sets, to confirm that our bounds reflect the behaviour of observed test error of QDA under these projections.

3 Results

Notation 1 *We denote scalars and vectors with lowercase letters and matrices with uppercase letters. The notations* $\mathbb{R}^n$ *and* $\mathbb{R}^{m\times n}$ *for the set of* n*-dimensional real vectors and the set of* $m \times n$ *real matrices respectively are standard. Also,* $\mathbb{S}^n_+$ *denotes the set of symmetric positive-semidefinite* $n \times n$ *matrices. For vectors,* $\|\cdot\|$ *denotes the Euclidean norm. For matrices,* $|\cdot|$ *denotes the determinant. We let* $\lambda_i(\cdot)$ *denote the* i*-th largest eigenvalue and* $\mathrm{tr}(\cdot)$ *denote the trace of a matrix. We also use* $\lambda_{\max}(\cdot)$ *and* $\lambda_{\min}(\cdot)$ *to denote the largest and smallest eigenvalues of a matrix respectively.*

Suppose the data distribution resides in $\mathbb{R}^d$ where the ambient dimension d is too large. To reduce dimension, consider a linear projection mapping $x \mapsto Mx$, where $M \in \mathbb{R}^{k\times d}$. Classification is then performed in $\mathbb{R}^k$, where $k \leq d$. We are interested in the effect of such dimensionality reduction on the Bayes error of QDA.

Recall the Bhattacharyya bound given in Theorem 1. First, the effect of dimensionality reduction on the parameters (means and covariances) is as follows. The class means will be mapped as

$$\mu_0 \mapsto M\mu_0 \text{ and } \mu_1 \mapsto M\mu_1 \tag{5}$$

and the class covariance matrices will be mapped as

$$\Sigma_0 \mapsto M\Sigma_0 M^T \text{ and } \Sigma_1 \mapsto M\Sigma_1 M^T. \tag{6}$$

Let us see what the Bhattacharyya bound becomes as a result of these mappings. To simplify notation, from this point on, let us write

$$\begin{aligned}\Sigma &:= \frac{\Sigma_0+\Sigma_1}{2} \text{ and } \delta_M \\ &:= \delta(M\mu_0, M\Sigma_0M^T, M\mu_1, M\Sigma_1M^T),\end{aligned} \tag{7}$$

where δ is the Bhattacharyya distance, defined in (4). Then, the bound of Theorem 1 becomes $\exp(-\delta_M)$, where

$$\begin{aligned}\delta_M := &\underbrace{\frac{1}{8}(\mu_0-\mu_1)^T M^T \left(M\Sigma M^T\right)^{-1} M(\mu_0-\mu_1)}_{:=E_1} \\ &+ \underbrace{\frac{1}{2}\log|M\Sigma M^T|}_{:=E_2} - \underbrace{\frac{1}{4}\log|M\Sigma_0 M^T|}_{:=E_3} \\ &- \underbrace{\frac{1}{4}\log|M\Sigma_1 M^T|}_{:=E_4}.\end{aligned} \tag{8}$$

This projected form of the Bhattacharyya bound is the object of our analysis. To upper bound $\exp(-\delta_M)$, we need to

lower bound δ_M that appears in (8). We do so, under the three projection methods we discussed, each of which corresponds to a different choice for the matrix M. These are the three main results of our paper, and we devote one subsection to each.

In what follows, we denote the ambient dimension of the original data space by d, and the projection dimension with k, where $k \leq d$. We only consider Euclidean vector spaces, so all norms and distances will be taken to be Euclidean. For completeness, we include in the appendix any supporting lemmas used for obtaining our results.

3.1 QDA under Gaussian random projection

The first dimensionality reduction method we consider is Gaussian random projection. Take a $k \times d$ matrix R filled with i.i.d. randomly sampled elements from $\mathcal{N}(0, 1)$, where $k \leq d$, and pre-multiply the d-dimensional data source to reduce dimension. While not a projection in a strict geometric sense, its name of random projection (RP) is established, due to the near-isometry property of such linear mappings. This projection method has thoroughly been studied, including its effects on LDA classification, as discussed in Sect. 2.

The Bhattacharyya bound under the mapping $x \mapsto Rx$ becomes $\exp(-\delta_R)$, where δ_R is given in (8). In this section, we are therefore interested in upper bounding $\exp(-\delta_R)$; we do this in the following theorem.

Theorem 2 (Bhattacharyya bound under random projection) *Let $R \in \mathbb{R}^{k\times d}$ be a matrix whose elements are sampled i.i.d. from $\mathcal{N}(0, 1)$. Then, the following is an upper bound for $\exp(-\delta_R)$,*

$$\exp\left(-\frac{(1-\epsilon)k\|\mu_0-\mu_1\|^2}{8(\sqrt{\mathrm{tr}(\Sigma)}+\sqrt{k\lambda_{\max}(\Sigma)}+\epsilon)^2}\right)\cdot \left(\frac{\sqrt{\sqrt{\mathrm{tr}(\Sigma_0)}+\sqrt{k\lambda_{\max}(\Sigma_0)}+\epsilon}\cdot\sqrt{\sqrt{\mathrm{tr}(\Sigma_1)}+\sqrt{k\lambda_{\max}(\Sigma_1)}+\epsilon}}{\left(\sqrt{\mathrm{tr}(\Sigma)}-\sqrt{k\lambda_{\max}(\Sigma)}-\epsilon\right)_+}\right)^k \tag{9}$$

with probability at least $1-2\exp(-\epsilon^2/2\lambda_{\max}(\Sigma))-\exp(-\epsilon^2/2\lambda_{\max}(\Sigma_0))-\exp(-\epsilon^2/2\lambda_{\max}(\Sigma_1))-\exp(-k\epsilon^2/4)$, where $(\cdot)_+ = \max\{\cdot, 0\}$.

From Theorem 2 we see that the error bound does not directly depend on the ambient dimension d but only through the traces of the covariance matrices relative to their largest eigenvalues in conjunction with the projection dimension k. This means that even though in the worst case the bound grows with d, whenever the data distribution does not fill the entire ambient space evenly, the error bound adapts to structure, essentially through a notion of intrinsic dimension. Indeed, the Bhattacharyya bound stays constant no matter how large d is, as long as these quantities, along with the distance between class centers stay the same. We will verify this experimentally in Sect. 4.

Furthermore, from Theorem 2 we see that our bounds tightens when traces of the individual class covariances Σ_0, Σ_1 get small and the trace of the pooled covariance Σ gets large. The trace of a covariance relative to its largest eigenvalue quantifies to what extent the distribution fills the ambient space (and termed as the effective rank [24])—hence we see that one instance where our bound is tightest is when the individual covariances have lowest rank while their sum has highest rank—that is exactly the structural trait assumed in [23]. However, in contrast to that approach, our bound holds without this assumption while it highlights and lets us read off such favourable traits.

Proof of Theorem 2 Since we need to upper bound $\exp(-\delta_R)$, we lower bound δ_R and take the exponential of the negative of the bound we derive. To this end, we have to lower or upper bound each of the four terms in δ_R separately, depending on their sign. For consistency of notations, we use the labels introduced in (8) for these four terms.

We lower bound E_1 with high probability as follows:

$$\begin{aligned}
E_1 &= \frac{1}{8}(\mu_0-\mu_1)^T R^T (R\Sigma R^T)^{-1} R(\mu_0-\mu_1) && (10)\\
&\geq \frac{1}{8}\lambda_{\min}((R\Sigma R^T)^{-1})\|R(\mu_0-\mu_1)\|^2 && (11)\\
&= \frac{\|R(\mu_0-\mu_1)\|^2}{8\lambda_{\max}(R\Sigma R^T)} && (12)\\
&\geq \frac{(1-\epsilon)k\|\mu_0-\mu_1\|^2}{8\lambda_{\max}(R\Sigma R^T)} && (13)\\
&\geq \frac{(1-\epsilon)k\|\mu_0-\mu_1\|^2}{8(\sqrt{\mathrm{tr}(\Sigma)}+\sqrt{k\lambda_{\max}(\Sigma)}+\epsilon)^2} && (14)
\end{aligned}$$

with probability at least $1-\exp(-k\epsilon^2/4)-\exp(-\epsilon^2/2\lambda_{\max}(\Sigma))$. We used the lower bound of Theorem 1 to obtain (11), Theorem 2 to obtain (13) and the upper bound of Theorem 4 to obtain (14).

Next, we lower bound E_2 with high probability:

$$\begin{aligned}
E_2 &= \frac{1}{2}\log|R\Sigma R^T| && (15)\\
&= \frac{1}{2}\log\prod_{i=1}^{k}\lambda_i(R\Sigma R^T) && (16)\\
&= \frac{1}{2}\sum_{i=1}^{k}\log\lambda_i(R\Sigma R^T) && (17)\\
&\geq \frac{k}{2}\log\lambda_{\min}(R\Sigma R^T) && (18)\\
&\geq \frac{k}{2}\log(\sqrt{\mathrm{tr}(\Sigma)}-\sqrt{k\lambda_{\max}(\Sigma)}-\epsilon)_+^2 && (19)\\
&= k\log(\sqrt{\mathrm{tr}(\Sigma)}-\sqrt{k\lambda_{\max}(\Sigma)}-\epsilon)_+ && (20)
\end{aligned}$$

with probability at least $1 - \exp(-\epsilon^2/2\lambda_{\max}(\Sigma))$. Here we used the lower bound given in Theorem 4 to obtain (19).

We upper bound E_3 with high probability:

$$E_3 = \frac{1}{4}\log|R\Sigma_0 R^T| \tag{21}$$

$$= \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(R\Sigma_0 R^T) \tag{22}$$

$$= \frac{1}{4}\sum_{i=1}^{k}\log\lambda_i(R\Sigma_0 R^T) \tag{23}$$

$$\leq \frac{k}{4}\log\lambda_{\max}(R\Sigma_0 R^T) \tag{24}$$

$$\leq \frac{k}{4}\log(\sqrt{\mathrm{tr}(\Sigma_0)} + \sqrt{k\lambda_{\max}(\Sigma_0)} + \epsilon)^2 \tag{25}$$

$$= \frac{k}{2}\log(\sqrt{\mathrm{tr}(\Sigma_0)} + \sqrt{k\lambda_{\max}(\Sigma_0)} + \epsilon) \tag{26}$$

with probability at least $1 - \exp(-\epsilon^2/2\lambda_{\max}(\Sigma_0))$. To obtain (25), we used the upper bound in Theorem 4.

Finally, we can upper bound E_4 in the same way as E_3, switching Σ_0 to Σ_1, which gives

$$E_4 \leq \frac{k}{2}\log(\sqrt{\mathrm{tr}(\Sigma_1)} + \sqrt{k\lambda_{\max}(\Sigma_1)} + \epsilon) \tag{27}$$

with probability at least $1 - \exp(-\epsilon^2/2\lambda_{\max}(\Sigma_1))$.

Combining the bounds on E_1, E_2, E_3 and E_4 by using the union bound, and plugging back into the exponential, we obtain the result stated in Theorem 2.

3.2 QDA under random orthogonal projection

Next, we consider random orthogonal projection to reduce dimensionality. This is similar to random projection but uses a random matrix that is semi-orthogonal[1] with a uniformly random orientation. It is expected that this will perform very similarly to the Gaussian random projection when d is large due to a well-known measure concentration effect by which the rows of R have nearly the same norms, and pairwise rows are near-orthogonal when d is large. However, R_o allows us to derive a slightly tighter bound.

To generate R_o, first we generate a $k \times d$ Gaussian RP matrix R as in Sect. 3.1, and then take $R_o := (RR^T)^{-1/2}R \in \mathbb{R}^{k\times d}$; this is indeed semi-orthogonal. Now, we are interested in upper bounding $\exp(-\delta_{R_o})$, which we do in the following theorem.

Theorem 3 (Bhattacharyya bound under random orthogonal projection) *Let $R \in \mathbb{R}^{k\times d}$ be a matrix whose elements are sampled i.i.d. from $\mathcal{N}(0,1)$ and define $\mathbb{R}^{k\times d} \ni R_o := (RR^T)^{-1/2}R$. Then, the following is an upper bound for $\exp(-\delta_{R_o})$,*

$$\exp\left(-\frac{(1-\epsilon)k\|\mu_0-\mu_1\|^2}{8(\sqrt{\mathrm{tr}(\Sigma)}+\sqrt{k\lambda_{\max}(\Sigma)}+\epsilon)^2}\right) \cdot\prod_{i=1}^{k}\left(\sqrt{\frac{\sqrt{\lambda_i(\Sigma_0)\lambda_i(\Sigma_1)}}{\lambda_{d-k+i}(\Sigma)}}\right) \tag{28}$$

with probability at least $1 - \exp(-\epsilon^2/2\lambda_{\max}(\Sigma)) - \exp(-k\epsilon^2/4)$.

The bound in Theorem 3 is tighter than the one we obtained in the case of Gaussian random projection, due to the random projection being orthogonal and thus offering more control over the eigenvalues of the projected covariance matrices, and indeed three of the four terms are bounded with probability 1.

Most importantly, as before in the case of Theorem 2, the ambient dimension d does not appear in the error bound of Theorem 3. The error depends on it only through the traces of the covariance matrices relative to their largest eigenvalues, displaying adaptivity to intrinsic dimension without the user needing to know its value.

Proof of Theorem 3 As in the proof of Theorem 2, we first lower bound δ_{R_o}, considering each term separately, and using the labels in (8).

For the term E_1, note that expanding out the definition of R_o gives

$$R_o^T(R_o\Sigma R_o^T)^{-1}R_o = R^T(R\Sigma R^T)^{-1}R. \tag{29}$$

Therefore, this term admits the same lower bound as we had in Sect. 3.1, namely

$$E_1 \geq \frac{(1-\epsilon)k\|\mu_0-\mu_1\|^2}{8(\sqrt{\mathrm{tr}(\Sigma)}+\sqrt{k\lambda_{\max}(\Sigma)}+\epsilon)^2} \tag{30}$$

with probability at least $1 - \exp(-k\epsilon^2/4) - \exp(-\epsilon^2/2\lambda_{\max}(\Sigma))$.

We lower bound E_2 using the lower side of Poincaré's separation theorem given in Theorem 3 to obtain

$$E_2 = \frac{1}{2}\log|R_o\Sigma R_o^T| \tag{31}$$

$$= \frac{1}{2}\log\prod_{i=1}^{k}\lambda_i(R_o\Sigma R_o^T) \tag{32}$$

$$\geq \frac{1}{2}\log\prod_{i=1}^{k}\lambda_{d-k+i}(\Sigma). \tag{33}$$

[1] Recall that a matrix $A \in \mathbb{R}^{m\times n}$, with $m \leq n$, is called semi-orthogonal, if and only if $AA^T = I_m$, where I_m is the $m \times m$ identity matrix.

To bound E_3, we use the upper side of Poincaré's separation theorem (Theorem 3),

$$E_3 = \frac{1}{4}\log|R_o\Sigma_0 R_o^T| \tag{34}$$
$$= \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(R_o\Sigma_0 R_o^T) \tag{35}$$
$$\leq \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(\Sigma_0). \tag{36}$$

Finally, we can upper bound E_4 in the same way as E_3, by switching Σ_0 to Σ_1,

$$E_4 \leq \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(\Sigma_1). \tag{37}$$

Plugging everything back into the exponential, and noting that the bounds on E_2, E_3, E_4 are deterministic, we obtain the result of Theorem 3.

3.3 QDA under principal component analysis

The third dimensionality reduction method we consider is PCA. We write the best rank-k approximation of Σ as

$$\tilde{\Sigma}_k = U_k^T\Lambda_k U_k, \tag{38}$$

where $U_k \in \mathbb{R}^{k\times d}$ is semi-orthogonal and $\Lambda_k \in \mathbb{S}_+^k$ is diagonal. The matrix U_k is our projection matrix in this case, so we have to upper bound $\exp(-\delta_{U_k})$. This is provided in the following theorem.

Theorem 4 (Bhattacharyya bound under PCA) *Let $U_k \in \mathbb{R}^{k\times d}$ be a semi-orthogonal matrix whose rows consist of the k principal eigenvectors of Σ, and $\tilde{\Sigma}_k$ be the best rank-k approximation of Σ. Then, the following is an upper bound for* $\exp(-\delta_{U_k})$,

$$\exp\left(-\frac{\|(\tilde{\Sigma}_k^+)^{1/2}(\mu_0-\mu_1)\|^2}{8}\right)\cdot\prod_{i=1}^{k}\left(\sqrt{\frac{\sqrt{\lambda_i(\Sigma_0)\lambda_i(\Sigma_1)}}{\lambda_i(\Sigma)}}\right). \tag{39}$$

where $\tilde{\Sigma}_k^+$ is the Moore-Penrose pseudoinverse of $\tilde{\Sigma}_k$.

Proof Unlike previously, here the dimensionality reducing transformation U_k is a deterministic function of Σ; however, as before, we aim to express the bound in terms of the parameters, so we work to eliminate isolated occurrences of U_k. As in the proofs of Theorems 2 and 3, we lower bound δ_{U_k}, considering each term separately and using the labels in (8).

The term E_1 is can be written as follows:

$$E_1 = \frac{1}{8}(\mu_0-\mu_1)^T U_k^T(U_k\Sigma U_k^T)^{-1}U_k(\mu_0-\mu_1) \tag{40}$$
$$= \frac{1}{8}(\mu_0-\mu_1)^T U_k^T\Lambda_k^{-1}U_k(\mu_0-\mu_1) \tag{41}$$
$$= \frac{1}{8}\|\Lambda_k^{-1/2}U_k(\mu_0-\mu_1)\|^2 \tag{42}$$
$$= \frac{1}{8}\|\Lambda_k^{-1/2}U_k(\mu_0-\mu_1)\|^2 \tag{43}$$
$$= \frac{1}{8}\|U_k^T\Lambda_k^{-1/2}U_k(\mu_0-\mu_1)\|^2 \tag{44}$$
$$= \frac{1}{8}\|(\tilde{\Sigma}_k^+)^{1/2}(\mu_0-\mu_1)\|^2 \tag{45}$$

where the semi-orthogonality of U_k was used to obtain (44).

The term E_2 can also be computed exactly as

$$E_2 = \frac{1}{2}\log|U_k\Sigma U_k^T| = \frac{1}{2}\log\prod_{i=1}^{k}\lambda_i(U_k\Sigma U_k^T)$$
$$= \frac{1}{2}\log\prod_{i=1}^{k}\lambda_i(\Sigma). \tag{46}$$

We upper bound E_3 as follows:

$$E_3 = \frac{1}{4}\log|U_k\Sigma_0 U_k^T| \tag{47}$$
$$= \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(U_k\Sigma_0 U_k^T) \tag{48}$$
$$\leq \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(\Sigma_0). \tag{49}$$

In the last line we used Theorem 3.

Swapping Σ_0 and Σ_1 in E_3 gives an upper bound on E_4:

$$E_4 \leq \frac{1}{4}\log\prod_{i=1}^{k}\lambda_i(\Sigma_1). \tag{50}$$

Combining the bounds of E_1, E_2, E_3 and E_4 and plugging back, we obtain the result of Theorem 4.

Note that, due to the deterministic nature of PCA (given the parameters), Theorem 4 holds with probability 1—unlike Theorems 2 and 3, which only hold with high probability.

As another observation, we note that the second term of Theorem 4 is almost the same as the second term of Theorem 3, the only difference being that the largest k eigenvalues of Σ appear in place of the smallest ones. This makes the bound for PCA strictly tighter than the bound for random orthogonal projection when the true means are equal.

Remark 1 PCA is classic, but there may be other viable alternatives to finding the rank-k approximation of Σ in conjunction with QDA. For instance, one could use the best rank-k approximation of either Σ_0 or Σ_1, or some convex combination thereof. Another natural choice would be to find the projection that maximises the Bhattacharyya distance under the projected parameters, that is, find

$$\arg\max_U \delta_U, \tag{51}$$

where $U \in \mathbb{R}^{k\times d}$ is semi-orthogonal. To maximize δ_U, we need to maximize E_1 and E_2 and minimize E_3 and E_4. Ignoring the term E_1 (assuming that it does not change significantly, or that the means are equal) we are left with maximizing the following three expressions

$$\frac{1}{2}\log|U\Sigma U^T|, \qquad -\frac{1}{4}\log|U\Sigma_0 U^T|, \text{ and}$$
$$-\frac{1}{4}\log|U\Sigma_1 U^T|. \tag{52}$$

Apart from the constants and the signs, the only major difference is the matrix that appears in these three expressions. This simultaneous optimisation problem appears to be not straightforward though, apart from the equal-means case, for which there exists a method using convex optimization [25]. In this context, in the general case, maximising only the first term from (52)—i.e. PCA—may then be interpreted or viewed as a simplification of the simultaneous optimisation problem (38).

Crucially, as in Theorems 2 and 3, again the ambient dimension does not appear in the bound for the PCA-QDA combination, which means that letting d grow indefinitely, while keeping other parameters unchanged, will prevent the bound from increasing. Interestingly, here the trace of covariances does not appear either, so it is less clear whether the error actually depends on a notion of ID.

4 Numerical experiments

We now proceed to testing experimentally the theoretical results we obtained, in conjunction with the out-of-sample test error. As in Sect. 3, we devote one subsection to each projection method. For each method, we are interested in the following questions:

1. How is the test error affected by the intrinsic dimension of the data distribution compared to the ambient dimension?
2. Under what conditions on the true parameters does the predictive performance withstand the arbitrary increase of the ambient dimension?

One should reckon that it would be rather difficult to design the appropriate experiments for studying these questions on a purely experimental basis. However, our theoretical analysis gives us insights about these questions already, which serves as a guide for gaining further insights from experimentation. We would like to highlight this as a nice example demonstrating the practical value of theoretical analysis, and indeed a test of it. A second sense in which we put our theoretical results to a test is by assessing to what extent the variations of our upper bounds, when varying certain parameters of the problem, agree with variations in empirical estimates of the predictive error. An upper bound only tells us that the error does not exceed a certain threshold with high probability, but this still permits a lot of variation below that threshold, which we assess through the forthcoming experiments.

Regarding our two research questions above, we have insights pertaining to the second question directly, while for the first one we can only conjecture a monotonic relation between the error and the ID prior to conducting the numerical work.

4.1 Technical preliminaries

Part of our experimental setup involves increasing the ambient dimension of the data, while keeping all other quantities that appear in the bounds fixed, most of which involve the eigenvalues of the covariance matrices Σ_0, Σ_1 and Σ. In this section we give the procedures necessary to achieve this.

The trace of a covariance relative to its largest eigenvalue is known as the *effective rank*. This is actually a notion of ID that reflects to what extent the distribution fills the ambient space.

Definition 1 (*Effective rank* [24]) Let $A \in \mathbb{S}_+^d$. The effective rank of A is defined as

$$r(A) := \frac{\operatorname{tr}(A)}{\lambda_{\max}(A)}. \tag{53}$$

It is straightforward to show that $1 \le r(A) \le \operatorname{rank}(A) \le d$, for all $A \in \mathbb{S}_+^d$. One advantage of the effective rank is that it is more detailed than the rank, as it considers the magnitude of all eigenvalues, rather than just counting the non-zero ones, and can also take on non-integer values. As we will soon see, controlling the parameters that appear in some of our bounds amounts to controlling the effective rank of a covariance matrix. Also, since it is a notion of ID, it makes sense to vary the effective rank (apart from the rank) of covariances to see its effects.

The second ingredient we need is the following result that allows us to control the largest eigenvalues of a matrix sum.

Theorem 5 *Let $A, B \in \mathbb{S}_+^d$ with orthogonal eigenvalue decompositions*

$$A = \sum_{i=1}^{d} \lambda_i(A) u_i u_i^T \text{ and } B = \sum_{i=1}^{d} \lambda_i(B) v_i v_i^T, \tag{54}$$

where $u_i, v_i \in \mathbb{R}^d$, for all $i \in \{1, \ldots, d\}$, and define $\mathbb{S}_+^d \ni C := A + B$. Assume further that for some m, such that $1 \leq m < d$, the following condition holds:

$$\text{span}\{u_1, \ldots, u_m\} = \text{span}\{v_1, \ldots, v_m\}. \tag{55}$$

Then, the m largest eigenvalues of C only depend on the m largest eigenvalues of A and B.

Proof Let $\mathcal{W}$ be the subspace spanned by the first m eigenvectors of A and B and $\mathcal{W}^\perp$ be its orthogonal complement. Note that $\mathcal{W}^\perp = \text{span}\{u_{m+1}, \ldots, u_d\} = \text{span}\{v_{m+1}, \ldots, v_d\}$. Thus, there exists an orthogonal matrix $V \in \mathbb{R}^{d \times d}$ that "separates" these two complements, that is we have

$$A = V \begin{bmatrix} A_0 & 0^T \\ 0 & A_1 \end{bmatrix} V^T \text{ and } B = V \begin{bmatrix} B_0 & 0^T \\ 0 & B_1 \end{bmatrix} V^T, \tag{56}$$

where $A_0, B_0 \in \mathbb{S}_+^m$ contain the m largest eigenvalues of A and B respectively, $A_1, B_1 \in \mathbb{S}_+^{d-m}$ contain the $d-m$ smallest eigenvalues of A and B respectively and $0 \in \mathbb{R}^{(d-m) \times m}$ is the zero matrix. Note that neither of A_0, A_1, B_0, B_1 is necessarily diagonal, although V can always be chosen to make one pair diagonal (either A_0 and A_1, or B_0 and B_1). The matrix C will therefore be

$$C = V \begin{bmatrix} A_0 + B_0 & 0^T \\ 0 & A_1 + B_1 \end{bmatrix} V^T. \tag{57}$$

For the eigenvalues, we can ignore V and V^T and only consider the block matrix in this case. Since $C \in \mathbb{S}_+^d$, we know that m of its eigenvalues lie in the upper-left block and $d-m$ lie in the lower-right block. To show that the largest m eigenvalues of C only depend on the largest m eigenvalues of A and B, we just need to show that they lie in the upper-left block. One way is to bound the quadratic forms of $A_0 + B_0$ and $A_1 + B_1$. To this end, it suffices to only consider vectors in $\mathcal{W}$ or $\mathcal{W}^\perp$. Let $x \in \mathcal{W}$ be a unit vector. We have

$$x^T(A_0 + B_0)x = x^T A_0 x + x^T B_0 x \geq \lambda_m(A) + \lambda_m(B). \tag{58}$$

Now let $y \in \mathcal{W}^\perp$ be a unit vector. We have

$$y^T(A_1 + B_1)y = y^T A_1 y + y^T B_1 y \leq \lambda_{m+1}(A) + \lambda_{m+1}(B). \tag{59}$$

Since $\lambda_m(A) + \lambda_m(B) \geq \lambda_{m+1}(A) + \lambda_{m+1}(B)$, the m largest eigenvalues of C lie in its upper-left block, whereas its $d-m$ smallest eigenvalues lie in its lower-right block. Thus, its m largest eigenvalues only depend on the m largest eigenvalues of A and B, as they lie in the block that contains the block matrices A_0 and B_0. This completes the proof.

Theorem 5 basically tells us that if we increase d by adding eigenvalues to A and B that are smaller than their respective m-th largest eigenvalues, and the condition in (55) continues to hold, then the m largest eigenvalues of their sum do not change. We will later give more details on how this result can be used in each experimental setup.

Remark 2 To obtain a matrix with a fixed rank or effective rank, we need to control its eigenvalues. While for the rank of a matrix this is straightforward, for the effective rank this is not always the case. To obtain a matrix with a fixed effective rank, we need control over both the sum of its eigenvalues and also its largest eigenvalue. While there might be more sophisticated methods of doing so, ours relies on a simple algebraic trick.

To sample the eigenvalues so that the effective rank equals r, we freely choose the largest eigenvalue λ_1, and then we sample $\lambda_2, \ldots, \lambda_d$ such that the following two conditions are satisfied:

$$\sum_{i=2}^{d} \lambda_i = (r-1)\lambda_1, \tag{60}$$

and

$$0 \leq \lambda_i \leq \lambda_1 \text{ for all } i \in \{2, \ldots, d\}, \tag{61}$$

where (60) ensures that the effective rank equals r, which can be confirmed by comparing it with Definition 1 after doing a little algebra, and (61) ensures that all eigenvalues are non-negative and that none of them is greater than λ_1. Clearly we must have $1 \leq r \leq d$ for these two conditions to hold simultaneously.

Sampling numbers such that they all fall in the same interval (in this case $[0, \lambda_1]$) and equal a fixed sum (in this case $(r-1)\lambda_1$) is a well-studied problem and routines exist in several programming languages to implement it (e.g. [26]). If one later needs to change the largest eigenvalue, while keeping the effective rank the same, they can just multiply all eigenvalues by the same positive constant. This is why the initial choice of λ_1 can be arbitrary.

After sampling the eigenvalues, we let the matrix in question take the following form

$$V \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d) V^T, \tag{62}$$

where $V \in \mathbb{R}^{d\times d}$ is an orthogonal matrix which is sampled randomly. The problem of sampling random orthogonal matrices has also been well-studied (e.g. [27]), and several routines exist in most programming languages. This setup makes the covariance matrices as random as possible, apart from a fixed effective rank. The form of (62) is also used to obtain a matrix with a fixed rank, after determining its eigenvalues.

4.2 Experimental setup

Each of the three projection methods we considered in the earlier sections has its own characteristics. Part of our setup is the same for all three, as we seek to answer our research questions and observe similarities and differences between the methods. We describe our setup here.

The first step is to fix the true parameters (means, covariances) of the two Gaussian classes, as well as the projection dimension. Then, we sample a test set in the ambient space from one of these. We then choose the projection matrix (randomly or deterministically, depending on the method) and project the test set using this matrix. Finally, we run QDA on the projected set using the projected means and covariances and record the empirical test error fraction measured on the test sample. Exploiting symmetry, we will use a test set sampled from class $\mathcal{C}_0$, and count the proportion of this test set that was classified to $\mathcal{C}_1$.

We repeat this strategy as we vary the values of the ambient and projection dimensions, and also repeat for different choices of the covariance matrices. In one type of experiments, we vary the rank or the effective rank of Σ_0 and Σ_1 together, that is, they will always have equal rank and effective rank. This reduces the number of experiments for each experimental setup. In another type of experiments, we vary the ambient dimension, while controlling the ranks and effective ranks of the covariance matrices. In both cases, we need to fix the largest eigenvalues of Σ_0 and Σ_1. We will set both of them to 1.

Equipped with these preliminaries and tools, we are now ready to present results obtained for each projection scheme separately. For completeness, we include the formal steps for all experiment designs we used in Appendix 2.

4.3 Numerical results with QDA under Gaussian RP

We start our experiments with Gaussian random projection. First we test how the empirical test error is affected by a low ID. To this end, we consider a setting where d remains fixed and the ID of Σ_0 and Σ_1 (measured by their rank and effective rank) is varied.

For each choice of the value of the ID, we randomly generate and fix the class mean parameters μ_0, μ_1 such that they have fixed distance of 1. We then generate and fix the class covariances Σ_0, Σ_1 using the procedure described in Remark 2. We also generate an out-of-sample test set of 1000 points.

We vary k, ensuring that it is always no higher than the rank (it can be higher than the effective rank). For each choice of k, we compute the test error. We report the average and standard deviation of the test errors from 100 independent random draws of the random projection matrix.

The complete steps of this set of experiments are given in Appendix 2 (Algorithm 1). The results are presented in Fig. 1. We see that, the lower the ID, in comparison with d, the lower the empirical test error. In addition, we also observe that, when we vary the rank, the empirical test error drops significantly faster than when we vary the effective rank.

In the next round of experiments we will vary the ambient dimension d. Our theoretical bound indicates that the error should not increase with d as long as other parameters in the bound, notably the effective ranks of covariance matrices, remain fixed. Our second set of experiments aims to test this behaviour on the actual empirical test errors. Therefore, we need to ensure that, as d increases, all other parameters that appear in Theorem 2 stay the same. Apart from k and ϵ, these are the following:

$$\|\mu_0 - \mu_1\|, \operatorname{tr}(\Sigma), \lambda_{\max}(\Sigma), \operatorname{tr}(\Sigma_0), \lambda_{\max}(\Sigma_0), \operatorname{tr}(\Sigma_1), \lambda_{\max}(\Sigma_1). \tag{63}$$

The mean distance $\|\mu_0 - \mu_1\|$ can be controlled by re-sampling the means at a fixed distance for each d. To fix $\lambda_{\max}(\Sigma_0)$ and $\lambda_{\max}(\Sigma_1)$, we just need to re-determine the covariance matrices such that their largest eigenvalues stay the same. After that, $\operatorname{tr}(\Sigma_0)$ and $\operatorname{tr}(\Sigma_1)$ can be fixed by keeping the effective ranks $r(\Sigma_0)$ and $r(\Sigma_1)$ the same for all d, according to its definition. We therefore re-sample the eigenvalues of Σ_0 and Σ_1 for each d, ensuring that they satisfy this property. Recall that Remark 2 described our method of how to do this.

The most difficult parameter to control is $\lambda_{\max}(\Sigma)$. This is because, unlike the trace, $\lambda_{\max}$ is not a linear function over $\mathbb{S}_+^d$. To control it, we use the result we established in Theorem 5 applied to $\Sigma_0/2$ and $\Sigma_1/2$ in place of A and B, and choosing $m = 1$. In this specific case, Theorem 5 says that, as d increases (and thus more eigenvalues are added), as long as the first principal eigenvectors of Σ_0 and Σ_1 remain collinear, the largest eigenvalue of Σ stays the same.

Having ensured the conditions of Theorem 5, we will create Σ_0 and Σ_1 to have the following form:

$$\Sigma_0 = \begin{bmatrix} \lambda_{\max}(\Sigma_0) & 0^T \\ 0 & D_0 \end{bmatrix} \text{ and } \Sigma_1 = \begin{bmatrix} \lambda_{\max}(\Sigma_1) & 0^T \\ 0 & V D_1 V^T \end{bmatrix}, \tag{64}$$

where $D_0 \in \mathbb{S}_+^{d-1}$ is diagonal with all eigenvalues of Σ_0 expect its largest, $D_1 \in \mathbb{S}_+^{d-1}$ is diagonal with all eigenvalues

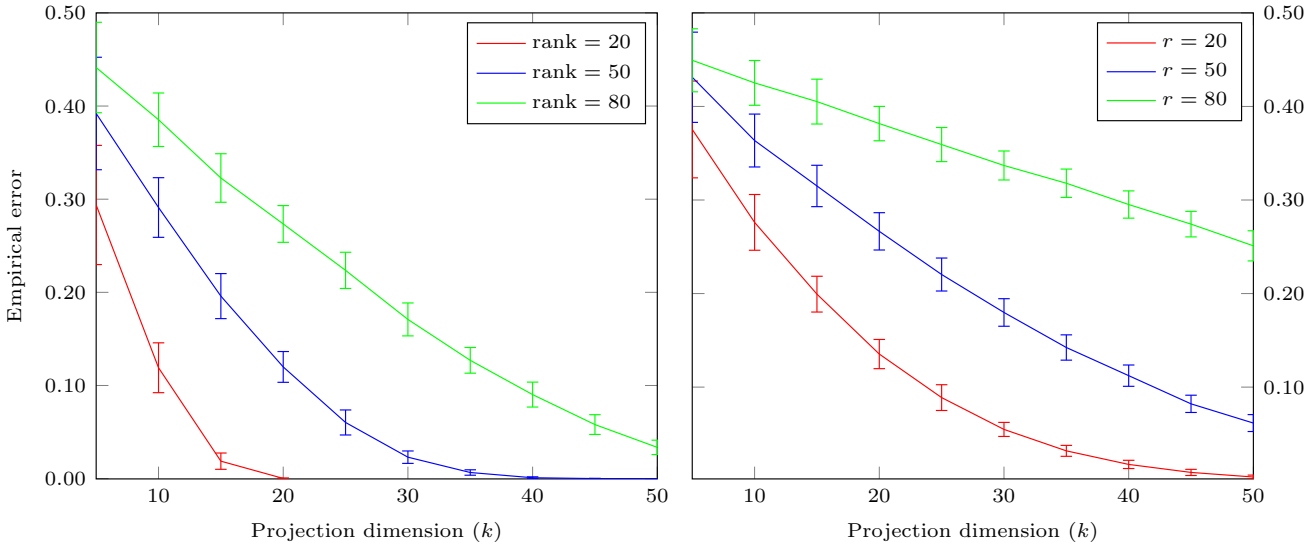

Fig. 1 Empirical out-of-sample error of QDA under Gaussian random projection when d is set to 100, and we vary the rank (left) or the effective rank (right) of both class covariance matrices while allowing them to span different subspaces. For each choice of rank or r, the mean vectors μ_0 and μ_1 have a fixed distance of 1 and Σ_0, Σ_1 are generated using the methodology of Remark 2, and then fixed. We show the mean empirical test error over 100 draws of the random matrix R for different values of k, each using a test set of size 1000 sampled i.i.d. from one of the Gaussian classes. The error bars span 1 standard deviation. The ambient dimension is set to 100

of Σ_1 expect its largest, $V \in \mathbb{R}^{(d-1)\times(d-1)}$ is orthogonal and $0 \in \mathbb{R}^{d-1}$ is the zero vector. The setup ensures that the eigenvectors corresponding to $\lambda_{\max}(\Sigma_0)$ and $\lambda_{\max}(\Sigma_1)$ both lie along the direction of the first axis, thus satisfying the condition of Theorem 5 for $m = 1$. The setup of (64) comes without loss of generality, and applying the same rotation to both Σ_0 and Σ_1 will result in the condition still holding, albeit for another direction.

As d increases, we re-sample the eigenvalues of Σ_0 and Σ_1, keeping both of their largest eigenvalues, as well as their effective ranks the same. The covariance matrices are then created anew using (64), which ensures that the conditions of Theorem 5 are still satisfied for $m = 1$. That is, the matrices D_0, D_1 and V grow in size but the first principal eigenvectors of Σ_0 and Σ_1 remain collinear.

The steps of this set of experiments are summarised in Appendix 2 (Algorithm 1). The results are given in Fig. 2. For all values of k, we see that, as d increases, the mean empirical test error fluctuates a little for small d, but then it stabilises for larger d of around 750. Most importantly, $k = 5$ which corresponds to an upper bound on the error in comparison to all larger values of k, the empirical test error stays almost constant for all values of d tested. This corroborates our theoretical finding that the error does not increase with d directly but only though the effective rank; in other words, compressive QDA is suitable in arbitrary high-dimensional problems as long as the effective rank is low.

The sudden drop of the empirical test error when d increases from 50 to 100 can be attributed to the fact that, since $r(\Sigma_0)$ and $r(\Sigma_1)$ remain fixed at 50, then at $d = 50$ the matrices are isotropic. Therefore, all of their eigenvalues are equal and the random projection results in a significant loss of information, regardless of the resulting low-dimensional subspace. When the dimension increases, however, some eigenvalues will have to decrease drastically to keep the effective rank fixed. Therefore, there are less directions of greatest variance, thus resulting in a much lower information loss, if the projection lands on a suitable subspace.

Finally, we remark that, intuitively, we would expect that the empirical test error will be higher when k is smaller. Our theoretical upper bound does not give us information on this, and it is therefore interesting to see from the empirical results in Fig. 2 that this is indeed the case in the examples tested.

We conclude from the results of this section that random projection is a suitable dimensionality reduction method for QDA when the underlying distribution has a low-ID covariance. The bound we derived for the generalisation error might be not particularly tight, but it captures most of the behaviour observed throughout our empirical investigations. In particular, the error does not grow with d but only through the effective rank.

4.4 Numerical results with QDA under random orthogonal projection

With random orthogonal projection-based dimensionality reduction we expect very similar results as those observed with Gaussian random projection. As argued in [9], and

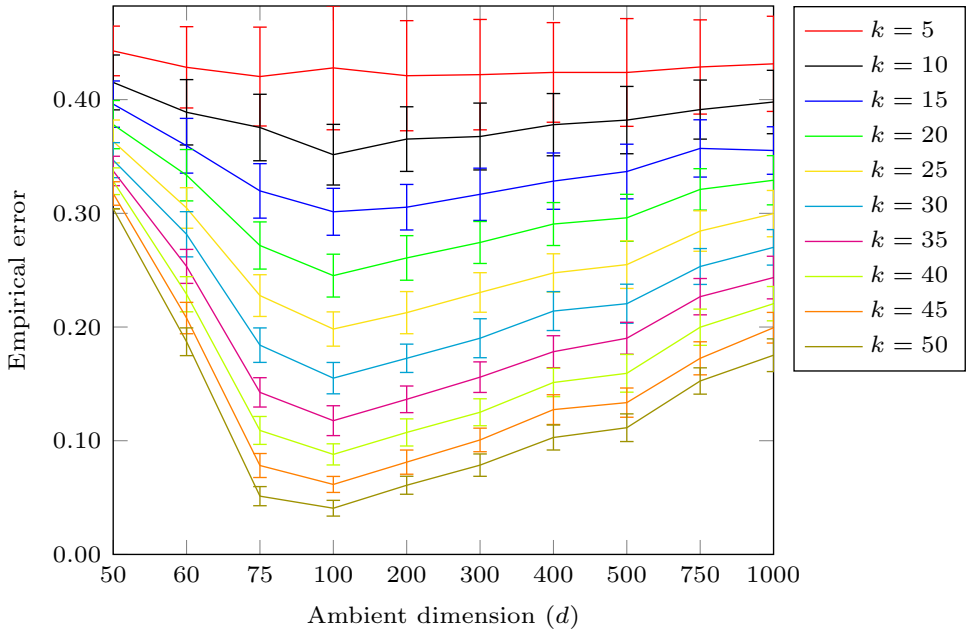

Fig. 2 Empirical test error of QDA under Gaussian random projection, when we vary d while keeping all quantities that appear in the bound of Theorem 2 fixed. For all d, $\|\mu_0 - \mu_1\| = 1$, $\lambda_{\max}(\Sigma_0) = \lambda_{\max}(\Sigma_1) = 1$, and $r(\Sigma_0) = r(\Sigma_1) = 50$. We show the average of the empirical test errors over 100 independent realizations of the random projection matrix R for different values of k, using a test set of size 1000 sampled i.i.d. from one of the Gaussian classes. The error bars span one standard deviation

proved formally in [28], when d is large, the columns of R with entries i.i.d. from $\mathcal{N}(0, 1)$, are close to being orthonormal.

The setup will be identical to the one used in random projection; for full reproducibility all steps are summarised in Appendix 2 (Algorithm 2). The first set of experiments aims to find out whether the empirical test error adapts to a low-ID, and the results are presented in Fig. 3.

As expected, the empirical test error is almost the same as in Fig. 1 in both of their respective subfigures, for all choices of ID and all values of k. This confirms experimentally that random projection essentially plays the role of random orthogonal projection, or equivalently, the matrix R used for random projection, is close to being semi-orthogonal. Therefore, not much difference is seen when orthonormalizing R.

For the second set of experiments, we test the extent to which the empirical test error does not depend on d. As before, we would like to create the class parameters in a way to ensure that all quantities that appear in the bound of Theorem 3 are fixed. While the first factor is the same as in Theorem 2, and can thus be controlled in the same way, for the second factor we would need to control the k largest eigenvalues of Σ_0 and Σ_1 and the k smallest eigenvalues of Σ. However, Theorem 5 only guarantees fixing the largest eigenvalues of Σ and we have not found a way to control its smallest.

Nevertheless, we expect random projection to behave very similarly to orthogonal projection, therefore we instead opt to use the same setup as we used there in Sect. 4.3, using the steps of Algorithm 2. The results are presented in Fig. 4.

As expected, the results appear nearly identical to those in Fig. 2. The conclusions are therefore the same as in Sect. 4.3, namely, increasing d does not blow up the errors as long as the quantities that appear in the bound stay unchanged.

As a final remark, we should mention that, despite the seemingly different bounds in Theorems 2 and 3, the empirical performance is very similar. Of course for smaller d, this might have not been the case, but the significance of our work is in the high-dimensional settings.

4.5 Numerical results with QDA under principal components analysis

We finally experiment with PCA-based dimensionality reduction for QDA. In this set of experiments, we again inspect the dependence of the empirical test error on the ID. The setup will be the same as in Sects. 4.3 and 4.4, except for the fact that, since the projection is deterministic given the parameters, there is no need to sample random matrices. We will therefore record the empirical test error directly for each pair (k, d). This reduces the required computational resources, albeit performing the eigen-decomposition is significantly more costly than a random projection. For reproducibility, all steps of this experiment are given in Algorithm 3, and the results are presented in Fig. 5.

From Fig. 5, we observe a completely different picture from the previously obtained Figs. 1 and 3. Here, when the rank is lower, the empirical test error seems to be unaffected by it. This is unlike random projection and random orthogonal projection, where the test error shows a clear dependence on the rank of the covariance matrices. However, for the effective rank, we see that a lower effective rank gives a lower empirical test error.

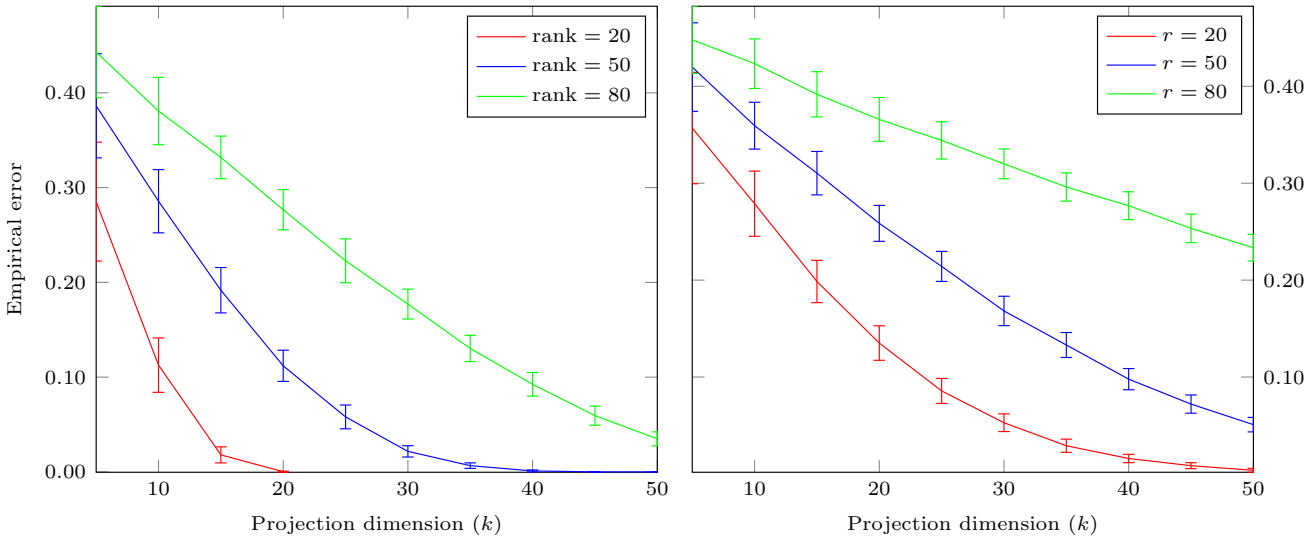

Fig. 3 Empirical out-of-sample error of QDA under random orthogonal projection when d is set to 100, and we vary the rank (left) or the effective rank (right) of both class covariance matrices while allowing them to span different subspaces. For each choice of rank or r, the mean vectors μ_0 and μ_1 have a fixed distance of 1 and Σ_0, Σ_1 are created using the methodology of Remark 2, and fixed. We show the mean empirical test error over 100 draws of the random orthonormal matrix R, for different values of k, each using a test set of size 1000 sampled i.i.d. from one of the Gaussian classes. The error bars span 1 standard deviation. The ambient dimension is set to 100

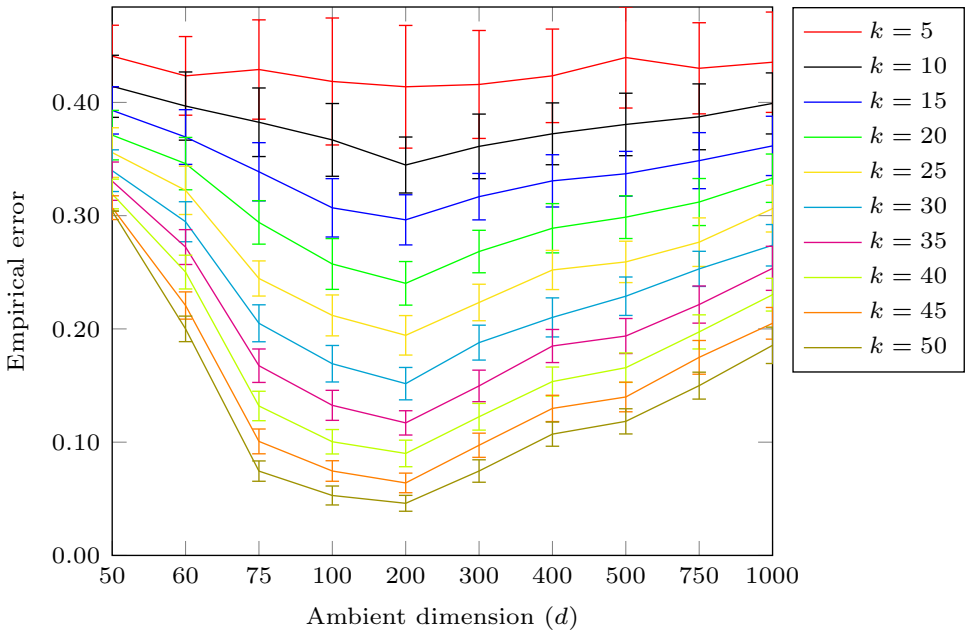

Fig. 4 Empirical test error of QDA under random orthogonal projection when varying d while we keep all quantities that appear in the bound of Theorem 3 fixed. For all d, we have $\|\mu_0 - \mu_1\| = 1$, $\lambda_{\max}(\Sigma_0) = \lambda_{\max}(\Sigma_1) = 1$, and $r(\Sigma_0) = r(\Sigma_1) = 50$. The condition of Theorem 5 is satisfied with $m = 1$. The solid lines show the average empirical test error over 100 realizations of the random matrix R for different values of k, using a test set of size 1000 sampled i.i.d. from only one Gaussian. The error bars span 1 standard deviation

For the second round, we need to fix all parameters in Theorem 4 as d increases. Apart from k, we need to control the following:

$$\|\Lambda_k^{-1/2} U_k(\mu_0 - \mu_1)\|, \{\lambda_i(\Sigma_0)\}_{i=1}^k, \{\lambda_i(\Sigma_1)\}_{i=1}^k, \{\lambda_i(\Sigma)\}_{i=1}^k. \tag{65}$$

To control the k largest eigenvalues of Σ_0, Σ_1 and Σ, Theorem 5 for $m \geq k$ provides a sufficient condition. Namely, we determine Σ_0 and Σ_1 to be of the following form:

$$\Sigma_0 = \begin{bmatrix} D_0 & 0^T \\ 0 & F_0 \end{bmatrix} \text{ and } \Sigma_1 = \begin{bmatrix} V D_1 V^T & 0^T \\ 0 & W F_1 W^T \end{bmatrix}, \tag{66}$$

where $D_0 \in \mathbb{S}_+^m$ and $F_0 \in \mathbb{S}_+^{d-m}$ are diagonal with the m largest and $d-m$ smallest eigenvalues of Σ_0 respectively, and similarly $D_1 \in \mathbb{S}_+^m$ and $F_1 \in \mathbb{S}_+^{d-m}$ for Σ_1, $V \in \mathbb{R}^{m \times m}$ and $W \in \mathbb{R}^{(d-m)\times(d-m)}$ are orthogonal and $0 \in \mathbb{R}^{(d-m)\times m}$ is the zero matrix. This is similar to the setting in (64), except that for each k we need the condition of Theorem 5 to hold with $m \geq k$. This is to ensure that, as d increases, only specified

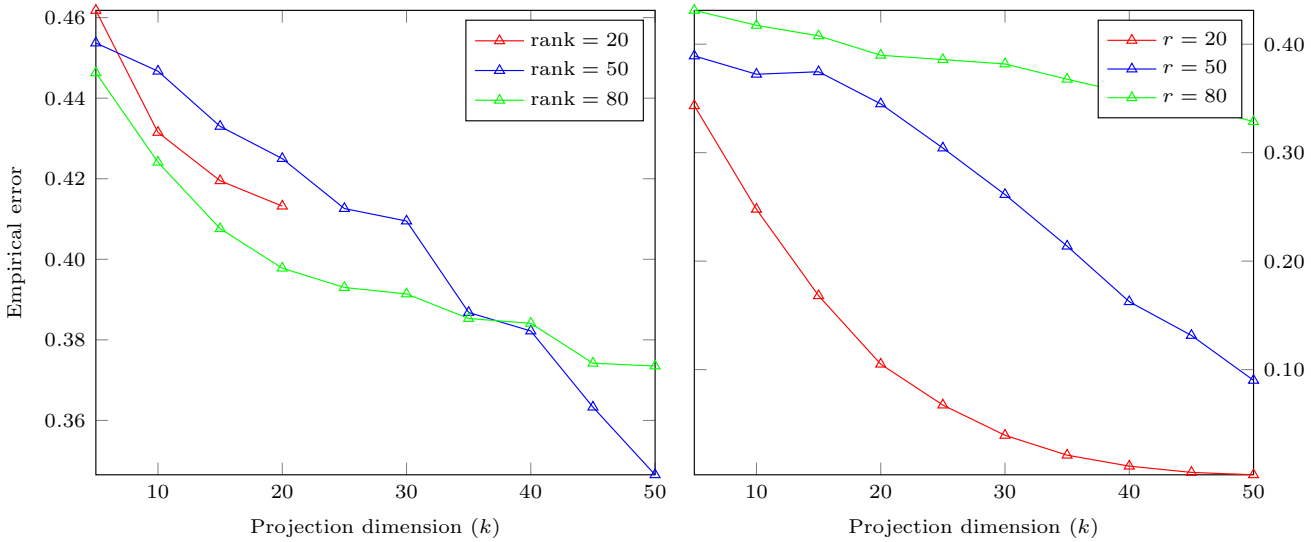

Fig. 5 empirical test error of QDA under PCA projection when we vary the rank (left) or the effective rank (right) of the covariance matrices. For each choice of rank and r, the class means have a distance of 1 and the covariance matrices were obtained using the methodology given in Remark 2. We show the empirical test error for different values of k, using a test set of size 10000 sampled i.i.d. from one of the Gaussian classes. The ambient dimension is set to 100

eigenvalues of Σ are considered. To show that $m \geq k$ is essential, we fix $m = 30$ in the experiments, so that we can expect to see some fluctuations in the empirical test error when $k > 30$.

Along with this condition, we need to fix the first k eigenvalues of the two covariance matrices. Therefore, we first create 50 eigenvalues, and then progressively add more as d increases, ensuring that all of them are smaller than the first 50. The covariance matrices are then created with (66). As d increases, the matrix V must remain fixed, in order to keep the upper-left block of Σ unchanged, whereas W grows in size and has to be re-determined.

To control the norm in (65), we opt to fix the vector appearing in the norm. To this end, we again require that the condition of Theorem 5 holds for $k \leq m$. This way, as d increases Λ_k remains the same and U_k obtains more columns that consist of zeros. Therefore, we just require that as d increases and more coordinates of $\mu_0 - \mu_1$ appear, its existing coordinates do not change. In other words, we determine all 1000 coordinates of μ_0 and μ_1 from the start and unveil d of them progressively as we increase d. Note that the equal-means case is just a specific case of this.

The complete sequence of steps for running these experiments is given in Algorithm 3. We present the results in Fig. 6. From this figure we can clearly see that the empirical test error remains very similar for all d, but only when $k \leq 30$. When $k > 30$, the empirical test error increases with d slightly, but it remains under an upper bound (e.g. that corresponds to $k = 5$). This is expected because the condition of Theorem 5 is satisfied for $m = 30$.

Comparing Figs. 2, 4 and 6, we observe that PCA tends to outperform both versions of random projection by about 15%, even when the means are closer to each other (since truncating coordinates reduces their distance for smaller d). The bound of PCA-QDA is also numerically tighter. These results also show the some precise conditions on the data, under which PCA outperforms the other two projection methods, i.e. the conditions of Theorem 5.

5 Conclusions

We derived upper bounds for the generalisation error of QDA when the data is subjected to three different alternative projection schemes. Specifically, we considered random Gaussian projection, random orthogonal projection and PCA. The first two make use of random matrices to reduce dimensionality, whereas the third one is a deterministic function of the model parameters. The bounds we derived are analysing the effect of these projections on the well-known Bhattacharyya bound. All three bounds turned out to be independent of the ambient dimension and instead depend on the effective rank – resembling an interesting dimension-adaptive behaviour recently found in the case of nuclear-norm regularised discriminative QDA [8]. We also confirmed our findings by extensive empirical simulations on synthetic data.

Future work can focus on deriving similar results for other dimensionality reduction methods (stochastic or deterministic) for QDA. The goal is to find the conditions under which the empirical test error adapts to a low-ID and the extend on which this is true.

Another possible research direction would be to examine whether other classifiers, apart from LDA and QDA, also

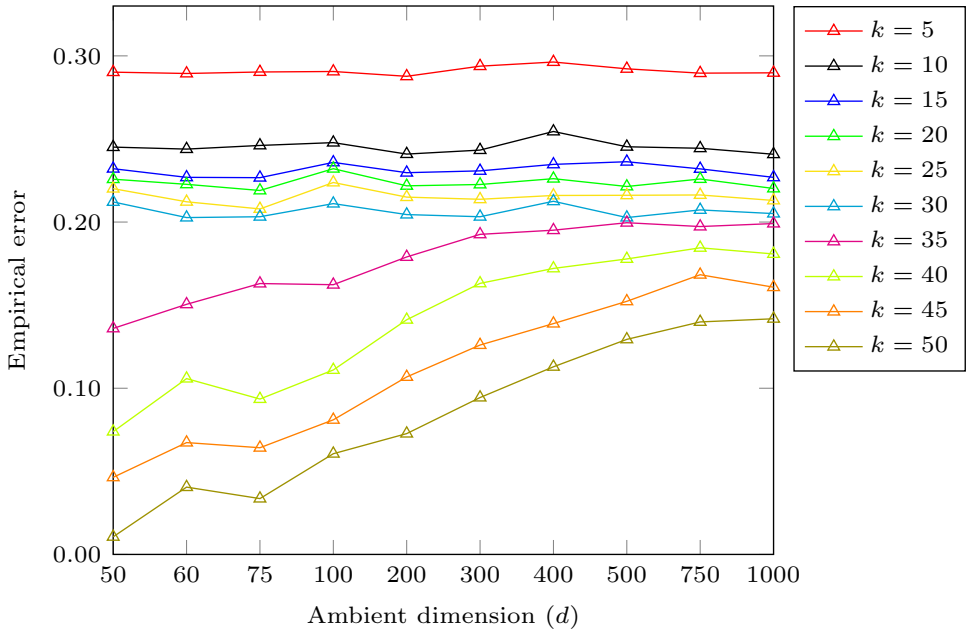

Fig. 6 Empirical test error of QDA under PCA when varying d while keeping all quantities that appear in Theorem 4 fixed. For all d, $\lambda_{\max}(\Sigma_0) = \lambda_{\max}(\Sigma_1) = 1$. The means were chosen such that $\|\mu_0 - \mu_1\| = 1$ when $d = 1000$ and are truncated for smaller d. The remaining eigenvalues of Σ_0 and Σ_1 were chosen such that $r(\Sigma_0) = r(\Sigma_1) = 50$ when $d = 1000$ and are truncated for smaller d. The condition of Theorem 5 is satisfied with $m = 30$. The lines show the empirical test error using a test set of size 10000 sampled i.i.d. from one of the Gaussian classes

enjoy a lower classification error in the presence of some low-ID assumption on the data generator, and to find out which notions of ID appear to capture this behaviour.

Acknowledgements AK was funded by EPSRC Fellowship EP/P004245/1 "FORGING: Fortuitous Geometries and Compressive Learning".

Author Contributions Conception & design: AK & EP; manuscript original draft: EP; revisions & editing: AK & EP; supervision: AK.

Funding EPSRC Fellowship EP/P004245/1.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest or competing interests relating to the content of this article.

Ethics approval This article does not contain any studies with human participants performed by any of the authors. This article does not contain any studies involving animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix A: Supplementary Lemmas

We include here any important lemmas we used to derive our new results in Sect. 3. Most of them are well-known inequalities. The reader is referred to the corresponding citations to find the proofs.

Lemma 1 (Rayleigh quotient [29]) *For any $A \in \mathbb{S}_+^d$ and non-zero $x \in \mathbb{R}^d$, we have*

$$\lambda_{\min}(A) \leq \frac{x^T A x}{\|x\|^2} \leq \lambda_{\max}(A). \tag{67}$$

Lemma 2 (Distributional Johnson-Lindenstrauss lemma [30]) *Given $0 < \epsilon < 1$, a fixed vector $x \in \mathbb{R}^d$ and a random matrix $R \in \mathbb{R}^{k \times d}$ whose elements are sampled i.i.d. from $\mathcal{N}(0, 1)$, then, with probability at least $1 - \exp(-k\epsilon^2/4)$, we have*

$$(1 - \epsilon)k\|x\|^2 \leq \|Rx\|^2. \tag{68}$$

Lemma 3 (Poincaré separation theorem [29]) *Let $A \in \mathbb{S}_+^d$ and let $B \in \mathbb{R}^{k \times d}$ be semi-orthogonal. Then, for all $i \in \{1, \ldots, k\}$, the following inequality holds:*

$$\lambda_{d-k+i}(A) \leq \lambda_i(BAB^T) \leq \lambda_i(A). \tag{69}$$

Lemma 4 (Eigenvalues of Gaussian random matrices [10]) *Let $R \in \mathbb{R}^{k \times d}$ be a random matrix whose elements are sampled i.i.d. from $\mathcal{N}(0, 1)$ and let $\Sigma \in \mathbb{S}_+^d$. Then, for all $\epsilon \geq 0$,*

with probability at least $1-\exp(-\epsilon^2/2\lambda_{max}(\Sigma))$*, we have*

$$(\sqrt{\mathrm{tr}(\Sigma)}-\sqrt{k\lambda_{\max}(\Sigma)}-\epsilon)_+^2 \le \lambda_{\min}(R\Sigma R^T), \tag{70}$$

provided that $k \le \lfloor \frac{\mathrm{tr}(\Sigma)}{\lambda_{\max}(\Sigma)} \rfloor$*, and with probability at least* $1-\exp(-\epsilon^2/2\lambda_{max}(\Sigma))$*, we have*

$$\lambda_{\max}(R\Sigma R^T) \le (\sqrt{\mathrm{tr}(\Sigma)}+\sqrt{k\lambda_{\max}(\Sigma)}+\epsilon)^2. \tag{71}$$

Appendix B: Experimentation details

We include here all steps involved in our experiments reported in Sect. 4.

Algorithm 1: empirical test error of QDA under random orthogonal projection.

Input: mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$; covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{S}_+^d$, projection dimension k; number of random matrices t; test set $T \subset \mathbb{R}^d$ of size n;

Output: $p_1, p_2, \ldots, p_t$.

for $i \in \{1, \ldots, t\}$ **do**
- Draw a random matrix $R \in \mathbb{R}^{k\times d}$ whose elements are i.i.d. from $\mathcal{N}(0,1)$;
- Set $p_i \leftarrow$ empirical test error on the set RT using QDA under the mapping $x \mapsto Rx$;

end for

Algorithm 2: empirical test error of QDA under random projection.

Input: mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$; covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{S}_+^d$, projection dimension k; number of random matrices t; test set $T \subset \mathbb{R}^d$ of size n;

Output: $p_1, p_2, \ldots, p_t$.

for $i \in \{1, \ldots, t\}$ **do**
- Draw a random matrix $R \in \mathbb{R}^{k\times d}$ whose elements are i.i.d. from $\mathcal{N}(0,1)$;
- Set $R_o \leftarrow (RR^T)^{-1/2}R$;
- Set $p_i \leftarrow$ empirical test error on the set R_oT using QDA under the mapping $x \mapsto R_ox$;

end for

Algorithm 3: empirical test error of QDA under PCA.

Input: mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$; covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{S}_+^d$, projection dimension k; number of random matrices t; test set $T \subset \mathbb{R}^d$ of size n;

Output: p.

Compute the orthogonal $d \times d$ matrix U that orthogonally diagonalizes $(\Sigma_0+\Sigma_1)/2$;

Set $U_k \leftarrow$ the k principal columns of U;

Set $p \leftarrow$ empirical test error on the set U_k^TT using QDA under the mapping $x \mapsto U_k^Tx$;

References

1. Verleysen, M., François, D.: The curse of dimensionality in data mining and time series prediction. In: International Work-conference on Artificial Neural Networks, pp. 758–770 (2005). Springer
2. Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D., Carin, L.: Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. IEEE Trans. Signal Process. **58**(12), 6140–6155 (2010)
3. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T.: The intrinsic dimension of images and its impact on learning. In: International Conference on Learning Representations (2020)
4. Kienitz, D., Komendantskaya, E., Lones, M.: The effect of manifold entanglement and intrinsic dimensionality on learning. Proc. AAAI Conf. Artif. Intell. **36**(7), 7160–7167 (2022)
5. Hamm, T., Steinwart, I.: Intrinsic dimension adaptive partitioning for kernel methods. SIAM J. Math. Data Sci. **4**(2), 721–749 (2022)
6. Suzuki, T., Nitanda, A.: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 3609–3621. Curran Associates, Inc. (2021)
7. Stubbemann, M., Hanika, T., Schneider, F.M.: Intrinsic Dimension for Large-Scale Geometric Learning (2022)
8. Latorre, F., Dadi, L.T., Rolland, P., Cevher, V.: The effect of the intrinsic dimension on the generalization of quadratic classifiers. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
9. Kabán, A., Durrant, R.J.: Dimension-adaptive bounds on compressive fld classification. In: International Conference on Algorithmic Learning Theory, pp. 294–308 (2013). Springer
10. Kabán, A.: Non-asymptotic analysis of compressive fisher discriminants in terms of the effective dimension. In: Asian Conference on Machine Learning, pp. 17–32 (2015). PMLR
11. Réfrégier, P., Galland, F.: Bhattacharyya bound for Raman spectrum classification with a couple of binary filters. Opt. Lett. **44**(9), 2228–2231 (2019)
12. Guillemot, V., Tenenhaus, A., Le Brusquet, L., Frouin, V.: Graph constrained discriminant analysis: a new method for the integration of a graph into a classification process. PLoS ONE **6**(10), 26146 (2011)
13. Shariatnia, S., Ziaratban, M., Rajabi, A., Salehi, A., Abdi Zarrini, K., Vakili, M.: Modeling the diagnosis of coronary artery disease by discriminant analysis and logistic regression: a cross-sectional study. BMC Med. Inform. Decis. Mak. **22**(1), 1–10 (2022)
14. Ilyasova, N.Y., Kupriyanov, A., Paringer, R.: The discriminant analysis application to refine the diagnostic features of blood vessels images. Opt. Memory Neural Netw. **24**(4), 309–313 (2015)
15. Li, M., Yuan, B.: 2d-lda: a statistical linear discriminant analysis for image matrix. Pattern Recogn. Lett. **26**(5), 527–532 (2005)
16. Guo, Y.-R., Bai, Y.-Q., Li, C.-N., Bai, L., Shao, Y.-H.: Two-dimensional Bhattacharyya bound linear discriminant analysis with its applications. Appl. Intell. **52**(8), 8793–8809 (2022)
17. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
18. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn (1990)
19. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, Hoboken (2000)
20. Ghojogh, B., Crowley, M.: Linear and quadratic discriminant analysis: Tutorial (2019). arXiv:1906.02590
21. Bhattacharyya, A.: On a measure of divergence between two multinomial populations. Sankhyā: Indian J. Stat. 401–406 (1943)
22. Kakade, S.M., Shalev-Shwartz, S., Tewari, A.: Regularization techniques for learning with matrices. J. Mach. Learn. Res. **13**(1), 1865–1890 (2012)
23. Reboredo, H., Renna, F., Calderbank, R., Rodrigues, M.R.: Bounds on the number of measurements for reliable compressive classification. IEEE Trans. Signal Process. **64**(22), 5778–5793 (2016)
24. Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science, vol. 47. Cambridge University Press (2020)
25. Cao, W.: Quadratic discriminant analysis revisited. City University of New York (2015)
26. Stafford, R.: Random Vectors with Fixed Sum (2022). https://www.mathworks.com/matlabcentral/fileexchange/9700-random-vectors-with-fixed-sum
27. Mezzadri, F.: How to generate random matrices from the classical compact groups. Not. Am. Math. Soc. **54**(5), 592–604 (2007)
28. Durrant, R.J., Kabán, A.: A tight bound on the performance of fisher's linear discriminant in randomly projected data spaces. Pattern Recogn. Lett. **33**(7), 911–919 (2012)
29. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2012)
30. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algor. **22**(1), 60–65 (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.