Dimension-Adaptive Bounds on Compressive FLD Classification

Ata Kabán and Robert J. Durrant

School of Computer Science, The University of Birmingham,
Birmingham, B15 2TT, UK

Abstract. Efficient dimensionality reduction by random projections (RP) gains popularity, hence the learning guarantees achievable in RP spaces are of great interest. In finite dimensional setting, it has been shown for the compressive Fisher Linear Discriminant (FLD) classifier that for good generalisation the required target dimension grows only as the log of the number of classes and is not adversely affected by the number of projected data points. However these bounds depend on the dimensionality d of the original data space. In this paper we give further guarantees that remove d from the bounds under certain conditions of regularity on the data density structure. In particular, if the data density does not fill the ambient space then the error of compressive FLD is independent of the ambient dimension and depends only on a notion of 'intrinsic dimension'.
Keywords: Random Projections, Compressed Learning, Intrinsic Dimension

1 Introduction and problem setting

A well known difficulty of machine learning in high dimensional data spaces is that the algorithms tend to require computational resources that grow exponentially with the data dimension. This is often referred to as the curse of dimensionality. Dimensionality reduction by random projections represents a computationally efficient yet theoretically principled way to alleviate this problem, and a new theory of learning based on this idea was already initiated in the work of [1]. Although the approach in [1] has some drawbacks, the idea to characterise learning in randomly projected data spaces has much unexplored potential.

More recent work in [5, 6] has analysed the performance of a compressive Fisher Linear Discriminant (FLD) classifier under assumption of full-rank covariance estimates, and has shown that its error rate with plug-in estimates can be upper-bounded in terms of quantities in the original data space, and the compressed dimensionality required for good generalisation grows only as the log of the number of classes. This result removed the number of projected points from the bounds, which was the main drawback in early approaches [1, 13] that relied on a global geometry preservation via the Johnson-Lindenstrauss lemma – however, perhaps unsurprisingly, the new bounds in [5, 6] now depend on the

dimensionality d of the original data space and the bounds get worse when d gets large. It is natural to ask how essential is this dependence?

Most often the high dimensional data does not fill the whole data space but exhibits some regularity. In such cases we would expect that learning should be easier [9]. A good theory of learning should reflect this. As noted in [9], an interesting question of great importance in itself is to identify algorithms whose performance scales with the 'intrinsic dimension' rather than the ambient dimension. For dimensionality reduction, this problem received a great deal of attention in e.g. subspace estimation and manifold learning [17, 10], but much less is known about dimension-adaptive generalisation guarantees [9] for e.g. classification or regression. Learning bounds for classification have mainly focused on data characteristics that hide dependence on the dimension, such as the margin. For randomly projected generic linear classifiers, a bound of the latter flavour has been recently given in [8]. In turn, here we seek guarantees in terms of a notion of 'intrinsic dimension' of the data space, and for this we focus on a specific classifier, the Fisher Linear Discriminant (FLD) working in a random subspace, which allows us to conduct a richer level of analysis.

1.1 Problem setting

We consider supervised classification, given a training set $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$ of N points where $(x_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}$ some (usually unknown) distribution on $Dom \times \mathcal{C}$ with the input domain Dom being R^d (in Section 2) or ℓ_2 more generally (in Section 3) and $y_i \in \mathcal{C}$, where $\mathcal{C}$ is a finite set of labels – e.g. $\mathcal{C} = \{0, 1\}$ for 2-class problems. For a given class of functions $\mathcal{F}$, the goal of learning a classifier is to learn from $\mathcal{T}_N$ the function $\hat{h} \in \mathcal{F}$ with the lowest generalisation error in terms of some loss function $\mathcal{L}$. That is, find $\hat{h} = \arg\min_{h \in \mathcal{F}} \mathrm{E}_{(x_q, y_q)}[\mathcal{L}(h)]$, where $(x_q, y_q) \sim \mathcal{D}$ is a random query point with unknown label y_q. We will use the $(0, 1)$-loss, which is most appropriate for 2-class classification, so we can write the generalisation error of a classifier $\hat{h} : Dom \to \{0, 1\}$ as

$$\mathrm{E}_{(x_q, y_q) \sim \mathcal{D}}[\mathcal{L}_{(0,1)}(\hat{h}(x_q), y_q) | \mathcal{T}_N] = \Pr_{(x_q, y_q)}[\hat{h}(x_q) \neq y_q | \mathcal{T}_N]$$

In this work the class of functions $\mathcal{F}$ will consist of Fisher Linear Discriminant (FLD) classifiers. We are interested in FLD that has access only to a randomly projected version of a fixed high dimensional training set, $\mathcal{T}_N^R = \{(Rx_i, y_i) : Rx_i \in \mathbb{R}^k, (x_i, y_i) \sim \mathcal{D}\}$ and we seek to bound the probability that a projected query point Rx_q is misclassified by the learnt classifier. This is referred to as the Compressive FLD.

FLD and Compressive FLD FLD is a simple and popular linear classifier, in widespread application. In its original form, the data classes are modelled as identical multivariate Gaussians, and the class label of a query point is predicted according to the smallest Mahalanobis distance from the class means. That is, denoting by $\hat{\Sigma}$ the empirical estimate of the pooled covariances and by $\hat{\mu}_0$ and

$\hat{\mu}_1$ the class mean estimates, the decision function of FLD at a query point x_q is:

$$\hat{h}(x_q) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^{-1}\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\}$$

where $\mathbf{1}(A)$ is the indicator function that returns one if A is true and zero otherwise. This can be derived from Bayes rule using the model of Gaussian classes $\mathcal{N}(\hat{\mu}_y, \hat{\Sigma})$ with equal weights.

Subjecting the data to a random projection (RP) means a linear transform by a $k \times d$ matrix R with entries drawn i.i.d. from $\mathcal{N}(0,1)$ (certain other random matrices are possible too). Although R is not a projection in strict mathematical sense, this terminology is widely established and it reflects the fact that when d is large the rows of a random matrix with i.i.d. entries are nearly orthogonal and have nearly equal lengths. The FLD estimated from a RP-ed training set will be denoted as $\hat{h}^R : \mathbb{R}^k \to \{0,1\}$, and this is:

$$\hat{h}^R(Rx_q) = \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\}$$

To facilitate analysis, the true distribution will also be assumed to consist of Gaussian classes as in classical texts [15], although it is clear from previous theoretical analyses [5, 6] that it is possible to relax this to the much wider class of sub-Gaussians. The true class means and covariances of these class-conditional densities will be denoted as μ_0, μ_1, Σ.

The generalisation error of $\hat{h}^R$, $\Pr_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q | \mathcal{T}_N, R]$, contains two independent sources of randomness: the training set $\mathcal{T}_N$, and the random projection R. Here we are interested to study how this quantity depends on the dimensionality of the data, and find conditions under which it exhibits dimension-adaptiveness. We start by writing the generalisation error of $\hat{h}^R$ to isolate the terms that affect its dependence on data dimension. We shall see that for a large enough sample size (of only $N > k + 2$) dimension adaptiveness is a property w.r.t. R, and it will be sufficient to study a simplified form of the error with the training set being kept fixed. To see this, decompose the generalisation error as in [7]: $\Pr_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q | \mathcal{T}_N, R] =$

$$= \sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2} \frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T R^T (R\hat{\Sigma}R^T)^{-1} R\,(\hat{\mu}_{\neg y} + \hat{\mu}_y - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\Sigma R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}}\right)$$

$$\leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-[E1 \cdot E2 - E3_y]\right) \tag{1}$$

where we used the Kantorovich and the Cauchy-Schwartz inequalities and defined:

$$E1 = \|(R\Sigma R^T)^{-\frac{1}{2}} R\,(\hat{\mu}_1 - \hat{\mu}_0)\,\| \tag{2}$$

$$E2 = \frac{\sqrt{\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})}}{1 + \kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})} \tag{3}$$

$$E3_y = \|(R\Sigma R^T)^{-\frac{1}{2}} R(\mu_y - \hat{\mu}_y)\| \tag{4}$$

and κ denotes condition number.

Now observe that $E2$ and $E3_y$ are estimation error terms in the k-dimensional projection space. Both of these can be bounded with high probability w.r.t. the random draws of $\mathcal{T}_N$, for any instance of R, in terms of k and N_0, N_1 and independent of R. Indeed, in the above[1], the contributions of both $E2$ and $E3_y$ vanish a.s. as N_0 and N_1 increase. In particular, for $N > k + 2$ the condition number in $E2$ (as a function of $\mathcal{T}_N$) is that of a Wishart $\mathcal{W}_k(N - 2, I_k)$, which is bounded w.h.p. [18] – even if N is not large enough for $\hat{\Sigma}$ to be full rank. Hence, these terms do not depend on the data dimension.

Furthermore, the norm of mean estimates that appears in $E1$ can be bounded from that of the true means independent of the ambient dimension also, using Lemma 1 in [7]. Therefore, to study the dimension-adaptiveness property of the error of compressive FLD it is sufficient to analyse the simplified 'estimated error' determined by $E1$ with $\mathcal{T}_N$ fixed, which we will denote as:

$$\hat{\Pr}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] = \Phi\left(-\frac{1}{2}E1\right) \tag{5}$$

Alternatively, we may study the limit of this quantity as $N0, N1 \to \infty$, which has the same form but with $\hat{\mu}_y$ replaced by μ_y (which is perhaps more meaningful to consider when we seek to show negative results by constructing lower bounds). This coincides with the Bayes error for the case of shared true class covariance, and will be denoted as $\Pr_{(x_q,y_q)}[h^R(Rx_q) \neq y_q]$. In the remainder of the paper we analyse these simplified error terms. We should note of course that for a complete non-asymptotic upper-bound on the generalisation error, the techniques in [7] may be used to include the contributions of all terms.

2 Some straightforward results in special cases

It is natural to ask if the error of compressive FLD could be bounded independently of the data dimension d. As we shall see shortly, without additional assumptions the answer is no in general. However, for data that exhibits some regularity in the sense that the data density does not fill the entire ambient space then this will be indeed possible. This section looks at three relatively straightforward cases for the sake of argument and insight.

2.1 Dependence on d cannot be eliminated in general

To start, we show that in general the dependence on d of the Compressed FLD error is essential. Assume Σ is full rank. We upper and lower bound the Bayes error to see that both bounds have the same dependence on d. First, notice that putting the orthonormalised $(RR^T)^{-1/2}R$ for R does not change eq.(1). Then

[1] Here we assumed equal class-conditional true covariances for convenience, although it is not substantially harder to allow these to differ while the model covariance $\hat{\Sigma}$ is shared.

using Rayleigh quotient ([11], Thm 4.2.2. pp. 176), the Poincaré inequality ([11], Corollary 4.3.16, pp. 190), and the Johnson-Lindenstrauss lemma [4] we get with probability at least $1 - 2\exp(\epsilon^2/4)$ the following:

$$\Pr_{(x_q,y_q)}[h^R(Rx_q) \neq y_q] \geqslant \Phi\left(-\frac{1}{2}\frac{\sqrt{(1+\epsilon)\cdot k}\cdot\|\mu_0-\mu_1)\|}{\sqrt{d\cdot\lambda_{\min}(\Sigma)}}\right) \tag{6}$$

$$\Pr_{(x_q,y_q)}[h^R(Rx_q) \neq y_q] \leqslant \Phi\left(-\frac{1}{2}\frac{\sqrt{(1-\epsilon)\cdot k}\cdot\|\mu_0-\mu_1)\|}{\sqrt{d\cdot\lambda_{\max}(\Sigma)}}\right) \tag{7}$$

Thus, it appears that a dependence on d of the generalisation error is the price to pay for not having required any 'sparsity-like' regularity of the data density. Figure 1 presents an empirical check that confirms this conclusion. In the next subsection we shall see a simple setting where such additional structure permits a better generalisation guarantee.

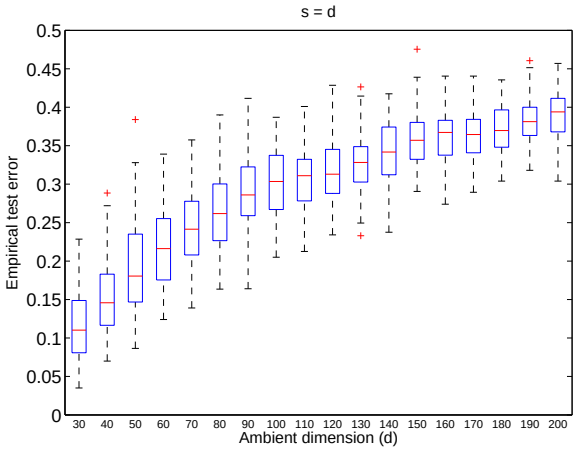

Fig. 1. Empirical error estimates of the compressive FLD as a function of the data dimension when the data does fill the ambient space and the distance between class centres stays constant. We see the error increases as we increase d. This confirms that the dependence of the error on d cannot be removed in general.

2.2 Case when the data density lives in a linear subspace

Consider the 2-class FLD, and $R \in \mathbb{R}^{k\times d}$ with entries from i.i.d. standard Gaussian, as before, but now consider the case when the entire data density lives in an s-dimensional linear subspace of the ambient space. We shall see, in this case the error can be upper-bounded in terms of s replacing d. This is formalised in the following result.

Theorem 1. *Let $(x_q, y_q) \sim \mathcal{D}$ a query point with unknown label y_q and Gaussian class conditional densities $x_q|_{y_q=y} \sim \mathcal{N}(\mu_y, \Sigma)$, and assume the distribution of the input points lives in an s-dimensional linear subspace of the ambient space $\mathbb{R}^d$. That is: $rank(\Sigma) = s < d$, and $\exists \boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \neq 0$ s.t. $\mu_0 = \mu_1 + \Sigma \boldsymbol{v}$. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with entries drawn i.i.d from $\mathcal{N}(0,1)$, with projection dimension $k \leqslant s$ (which is the case of interest for compression). Then, with probability at least $1 - \exp(-k\epsilon^2/4)$ over the random choice of R, $\forall \epsilon \in (0,1)$, we have the following:*

$$\hat{Pr}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] \leqslant \Phi\left(-\frac{1}{2}\frac{\sqrt{1-\epsilon}}{\sqrt{s}}\frac{\sqrt{k}\cdot\|\hat{\mu}_0 - \hat{\mu}_1\|}{\sqrt{\lambda_{\max}(\Sigma)}}\right)$$

Proof. By the low rank precondition, Σ equals its rank-s SVD decomposition, so we write $\Sigma = PSP^T$, where S contains an $s \times s$ full-rank diagonal matrix and zeros everywhere else, and $P \in \mathbb{R}^{d\times d}, P^TP = I = PP^T$. Replacing this into eq. (5) gives:

$$\hat{\mathrm{Pr}}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] = \Phi\left(-\frac{1}{2}\sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T R^T [RPSP^T R^T]^{-1} R(\hat{\mu}_0 - \hat{\mu}_1)}\right) \tag{8}$$

Next, observe that by construction $P^T(\hat{\mu}_0 - \hat{\mu}_1) = \hat{\mu}_0 - \hat{\mu}_1$ (since $\mu_0 - \mu_1 \in Range(\Sigma)$) and so $\hat{\mu}_0 - \hat{\mu}_1 \in Range(\Sigma)$ also).

Using these and denoting $\bar{R} = RP$,

$$(\hat{\mu}_0 - \hat{\mu}_1)^T R^T (R\Sigma R^T)^{-1} R(\hat{\mu}_0 - \hat{\mu}_1) \tag{9}$$

has the same distribution as:

$$(\hat{\mu}_0 - \hat{\mu}_1)^T P\bar{R}^T (\bar{R}S\bar{R}^T)^{-1} \bar{R}P^T(\hat{\mu}_0 - \hat{\mu}_1) \tag{10}$$

where, by the rotation-invariance of Gaussians, $\bar{R}$ is a $k \times s$ random matrix with i.i.d. standard normal entries.

Now, let $\bar{R}_o = (\bar{R}\bar{R}^T)^{-1/2}\bar{R}$. We can equivalently rewrite eq.(10), and then bound it as the following:

$$\begin{aligned} &= (\hat{\mu}_0 - \hat{\mu}_1)^T P\bar{R}_o^T (\bar{R}_o S\bar{R}_o^T)^{-1} \bar{R}_o P^T(\hat{\mu}_0 - \hat{\mu}_1) \\ &\geqslant \frac{\|\bar{R}_o P^T(\hat{\mu}_0 - \hat{\mu}_1)\|^2}{\lambda_{\max}(\bar{R}_o S\bar{R}_o)} \geqslant \frac{\|\bar{R}_o P^T(\hat{\mu}_0 - \hat{\mu}_1)\|^2}{\lambda_{\max}(S)} \qquad (11) \\ &= \frac{\|\bar{R}_o P^T(\hat{\mu}_0 - \hat{\mu}_1)\|^2}{\lambda_{\max}(\Sigma)} \qquad (12) \end{aligned}$$

where in the last two steps we used minorisation by Rayleigh quotient and the Poincaré inequality respectively — note that the latter requires R_o to be orthonormal.

Finally, we bound eq. (12) by Johnson-Lindenstrauss lemma [4], so $\|\bar{R}_o(P^T\hat{\mu}_0 - P^T\hat{\mu}_1)\|^2 \geqslant (1-\epsilon)\cdot k/s \cdot \|P^T\hat{\mu}_0 - P^T\hat{\mu}_1\|^2$ w.p. $1 - \exp(-k\epsilon^2/4)$, and use again that $\|P^T\hat{\mu}_0 - P^T\hat{\mu}_1\|^2 = \|\hat{\mu}_0 - \hat{\mu}_1\|^2$ to conclude the proof. □

Figure 2 presents an illustration and empirical validation of the findings of Theorem 1 employing synthetic data with two 5-separated Gaussian classes that live in $s < d = 100$ dimensions.

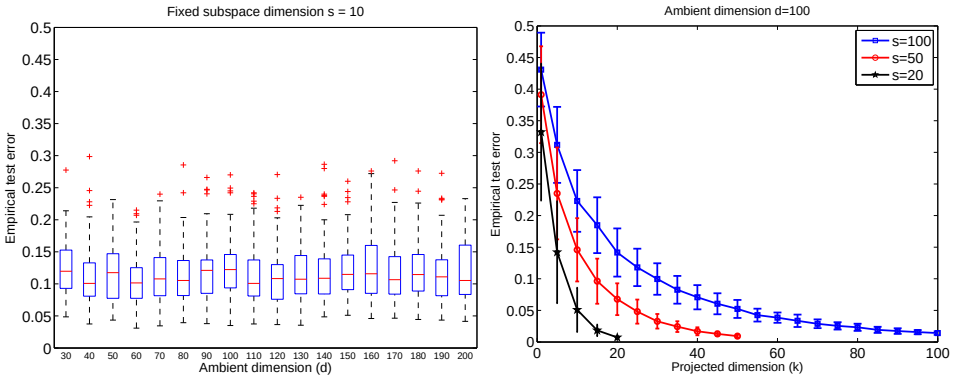

Fig. 2. Empirical performance when data density lives in a subspace. *Left*: When the data lives in a fixed subspace, then increasing the ambient dimension leaves the error constant. *Right*: With fixed ambient dimension ($d = 100$), a smaller dimension of the subspace where the data density lives implies a lower misclassification error rate of RP-FLD.

2.3 Noisy subspace

Now consider the case when the data density lives 'mostly' on a subspace up to some additive noise. We can show in this case that again the error may depend on d in general. To see this let us take $\Sigma = PSP^T + \sigma^2 I$ where S is an $s \times s$ full rank matrix embedded by P in the ambient space $\mathbb{R}^d$. We have:

$$\Pr_{(x_q,y_q)}[h^R(Rx_q) \neq y_q] = \Phi\left(-\frac{1}{2}\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T [R(\Sigma + \sigma^2 I)R^T]^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}\right)$$

and we lower and upper bound this.

Using Johnson-Lindenstrauss [4] and the Weyl's inequality, this can be lower-bounded as:

$$\geqslant \Phi\left(-\frac{1}{2}\sqrt{\frac{k(1+\epsilon)\|\mu_1 - \mu_0\|^2}{\lambda_{\min}(RP^T SPR^T) + \sigma^2 \lambda_{\min}(RR^T)}}\right)$$

$$\geqslant \Phi\left(-\frac{1}{2}\sqrt{\frac{k(1+\epsilon)\|\mu_1 - \mu_0\|^2}{\lambda_{\min}(S)(\sqrt{s} - \sqrt{k} - \nu)^2 + \sigma^2(\sqrt{d} - \sqrt{k} - \nu)^2}}\right)$$

w.p. $1 - \exp(-k\epsilon^2/4) - 2\exp(-\nu^2/2)$, $\forall \nu > 0, \forall \epsilon \in (0,1)$. In the last step we used Eq. (2.3) in [18] that lower-bounds the smallest singular value of a Gaussian random matrix.

Likewise, the same can be also upper-bounded using similar steps and the corresponding bound on the largest singular values [18], yielding:

$$\Pr_{(x_q,y_q)}[h^R(Rx_q) \neq y_q] \leqslant \Phi\left(-\frac{1}{2}\sqrt{\frac{k(1-\epsilon)\|\mu_1 - \mu_0\|^2}{\lambda_{\max}(S)(\sqrt{s} + \sqrt{k} + \nu)^2 + \sigma^2(\sqrt{d} + \sqrt{k} + \nu)^2}}\right)$$

w.p. $1 - \exp(-k\epsilon^2/4) - 2\exp(-\nu^2/2)$, $\forall \nu > 0, \forall \epsilon \in (0,1)$.

We see that both bounds depend on d at the same rate. So again, such a bound becomes less useful when d is very large unless either the separation of means $\|\mu_1 - \mu_0\|$ grows with d at least as $\sigma\sqrt{d}$, or the noise variance σ^2 shrinks as $1/d$. In the next section we consider data spaces that are separable Hilbert spaces (so $\|\mu_1 - \mu_0\|$ is finite whereas d can be infinite) equipped with a Gaussian measure, and we give conditions that ensure that the error remains bounded.

3 Main result: A Bound on Compressive Functional FLD

In this section the data space is a separable Hilbert space of possibly infinite dimension, here taken to be ℓ_2, equipped with Gaussian probability measure over Borel sets [14, 2], and we require that the covariance operator is trace class – i.e. its trace must be finite. As we shall see, this requirement ensures that the error of Compressive FLD can be bounded independent of the ambient dimension.

Definition [18]. The effective rank of Σ is defined as $r(\Sigma) = \frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}$.

The following main result provides a bound on the error of functional FLD that operates in a random k-dimensional subspace of the data space ℓ_2. This bound is in terms of the effective rank of Σ, which may be thought of as a notion of the intrinsic dimension of the data. The case of interest for compression is when k is small, and we will assume that $k \leqslant C \cdot r(\Sigma)$ for some constant $C > 0$ – as an analogue to the case $k \leqslant d$ typically taken in finite d settings.

Theorem 2. *Let $(x_q, y_q) \sim \mathcal{D}$ a query point with unknown label y_q and Gaussian class conditionals $x_q|_{y_q=y} \sim \mathcal{N}(\mu_y, \Sigma)$, where Σ is a trace-class covariance (i.e. $Tr(\Sigma_y) < \infty$); let $\pi_y = Pr(y_q = y)$, and let m be the number of classes. Let $(R_{1,i})_{i\geqslant 1}, \cdots (R_{k,i})_{i\geqslant 1}$ be k infinite sequences of i.i.d. standard normal variables, and denote by R the matrix whose rows are these sequences. For random projections from $\mathcal{H}$ onto $\mathbb{R}^k$ with k with $k \leqslant C \cdot r(\Sigma)$ for some positive constant C, we have that, $\forall \epsilon \in (0,1)$, $\forall \eta \in \left(0, \frac{\sqrt{k/r(\Sigma)}}{1+2\sqrt{\log 5}\cdot\sqrt{C}}\right]$, the error is bounded as the following:*
a) In 2-class case ($m = 2$), we have:

$$\hat{Pr}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] \leqslant \Phi\left(-\frac{1}{2}\frac{\sqrt{(1-\epsilon)k}\|\hat{\mu}_0 - \hat{\mu}_1\|}{\sqrt{Tr(\Sigma)}\left(1 + 4\sqrt{C\log(1+2/\eta)}\right)}\right) \tag{13}$$

with probability at least $1 - (\exp(-k\epsilon^2/4) + \exp(-k\log(1+2/\eta))$.
b) In multi-class case ($m > 2$), we have:

$$\hat{Pr}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] \leqslant \sum_{y=0}^{m-1}\pi_y \sum_{i=0,i\neq y}^{m-1} \Phi\left(-\frac{1}{2}\frac{\sqrt{(1-\epsilon)k}\|\hat{\mu}_y - \hat{\mu}_i\|}{\sqrt{Tr(\Sigma)}\left(1 + 4\sqrt{C\log(1+2/\eta)}\right)}\right) \tag{14}$$

with probability at least $1 - (\frac{m(m-1)}{2}\exp(-k\epsilon^2/4) + \exp(-k\log(1+2/\eta))$.

Now, looking at eq. (13) of Theorem 2 and its finite dimensional analogue in Theorem 1 (in the case of shared Σ) comparatively, we see the essential difference is that s is now replaced by $r(\Sigma)\left(1+4\sqrt{C\log(1+2/\eta)}\right)^2$, i.e. a small multiple of our notion of intrinsic dimension in ℓ_2.

The proof will make use of covering arguments. It is likely that the logarithmic factor $\log(1+2/\eta)$ could be removed with the use of more sophisticated proof techniques, however we have not pursued this here. Section 3.2 will give the details of the proof of Theorem 2.

An important consequence of this result is that despite the infinite dimensional data space, the order of the required dimensionality of the random subspace is surprisingly low – this is discussed in the next subsection.

3.1 Dimension of the compressive space

The projection dimension k required for good generalisation may be thought of as a measure of the difficulty of the task. It is desirable for a theory of learning to provide guarantees that reflect this. Early attempts to create RP learning bounds based on the strong global guarantees offered by the Johnson-Lindenstrauss lemma, e.g. [1] fell short of this aim and yielded a dependence of the order $k = \mathcal{O}(\log N)$ – where N is the number of training points that get randomly projected. A sharp improvement, under full covariance assumptions in fixed finite dimensions, [5] has shown that k only needs to be of the order $\mathcal{O}(\log m)$ for good classification guarantees, and this matches earlier results for unsupervised learning of a mixture of Gaussians [3].

However, because the ambient dimension d was a constant in these works, the previous bounds are not directly applicable when d is allowed to be infinity. In turn, we can now obtain as a consequence of Theorem 2 that under its conditions the required projection dimension for m-class classification is still $\mathcal{O}(\log m)$ independently of d:

Corollary 1. *With the notations and preconditions of Theorem 2, in order that the probability of misclassification for an m-class problem in the projected space remains below any given δ it is sufficient to take:*

$$k = \mathcal{O}(\log m)$$

Proof. The r.h.s. of part b) in Theorem 2 can be upper-bounded using Eq (13.48) of [12] for $\Phi(\cdot)$:

$$\leqslant \frac{1}{2}\sum_{y=0}^{m-1}\pi_y \sum_{i=0;i\neq y}^{m-1} \exp\left(-\frac{1}{8}\frac{(1-\epsilon)k\|\hat{\mu}_y-\hat{\mu}_i\|^2}{Tr(\Sigma)\left(1+4\sqrt{C\log(1+2/\eta)}\right)^2}\right)$$

Setting this to some $\delta \in (0,1)$ gives:

$$\log\left(\frac{m-1}{2\delta}\right) \leqslant \frac{1}{8}\frac{(1-\epsilon)\cdot k\cdot \min_{i,j=1,\ldots,m;i\neq j}\|\hat{\mu}_i-\hat{\mu}_j\|^2}{Tr(\Sigma)\left(1+4\sqrt{C\log(1+2/\eta)}\right)^2}$$

where we used that $\sum_{y=0}^{m-1} \pi_y = 1$. Solving for k we obtain

$$k \geqslant 8 \cdot \frac{Tr(\Sigma)(1+4\sqrt{C\log(1+2/\eta)})^2}{(1-\epsilon)\min_{i,j=0,\ldots,m-1, i\neq j} \|\hat{\mu}_i - \hat{\mu}_j\|^2} \cdot \log\left(\frac{m-1}{2\delta}\right) = \mathcal{O}(\log m) \tag{15}$$

Finally, for $k = \mathcal{O}(\log m)$ it is easy to see that the probability with which the bound holds in Theorem 2 part b) can be made arbitrarily small. □

Comparing the bound in eq. (15) with Corollary 4.10 in [5], we see that $d \cdot \lambda_{\max}(\Sigma)$ is now replaced by $Tr(\Sigma)(1+4\sqrt{C\log(1+2/\eta)})^2$ and may indeed be interpreted as the 'diameter' of the data that now depends only on the intrinsic dimension, while $\min_{i\neq j} \|\mu_i - \mu_j\|$ in the bound remains an analogue of the 'margin'.

Application One context in which functional data spaces are of interest is kernel methods. By way of demonstration, we conduct experiments with kernel-FLD (KFLD) restricted to a random k-dimensional subspace of the feature space. This is equivalent with a random compression of the gram matrix. Our bound in Theorem 2 applies to this case too, since the orthogonal projection of Σ into the span of the training points (i.e. the feature space) can only decrease the trace. We use 13 UCI benchmark datasets from [16], together with their experimental protocol. These data are: diabetes (N=468), ringnorm (N=400), waveform (N=400), flare solar (N=666), german (N=700), thyroid (N=140), heart (N=170), titanic (N=150), breast cancer (N=200), twonorm (N=400), banana (N=400), image (1300), splice (N=1000). Figure 3 summarises the results obtained for various choices of k and we see indeed that small values of k already produce results that are comparable to the full KFLD.

3.2 Proof of Theorem 2

The main ingredient of the proof is a bound on the largest eigenvalue of the projected covariance operator $R\Sigma R^T$, which is a corollary of the following theorem.

Theorem 3. *Let Σ a covariance operator s.t. $Tr(\Sigma) < \infty$ in a Gaussian Hilbert space $\mathcal{H}$ (assumed w.l.o.g. to be infinite dimensional), and let $(R_{1,i})_{i\geqslant 1}, \cdots (R_{k,i})_{i\geqslant 1}$ be k sequences of i.i.d. standard normal variables. Then, $\forall \eta \in (0,1)$, we have with probability at least $1 - \exp(-k\log(1+2/\eta))$:*

$$\lambda_{\max}(R\Sigma R^T) \leqslant \frac{Tr(\Sigma)}{(1-\eta)^2}\left(1 + 2\sqrt{\frac{k\cdot\lambda_{\max}(\Sigma)}{Tr(\Sigma)}\log(1+2/\eta)}\right)^2 \tag{16}$$

Proof of Theorem 3. Let us denote the unit sphere in $\mathbb{R}^k$ by $\mathcal{S}^{k-1}$. We use the covering technique on the sphere in three steps as follows.

Step 1 [Concentration] Let $\mathbf{w} \in \mathcal{S}^{k-1}$ fixed. Then, $\forall \epsilon > 0$,

$$\frac{\|\Sigma^{1/2}R^T\mathbf{w}\|^2}{Tr(\Sigma)} \leqslant 1 + \epsilon \tag{17}$$

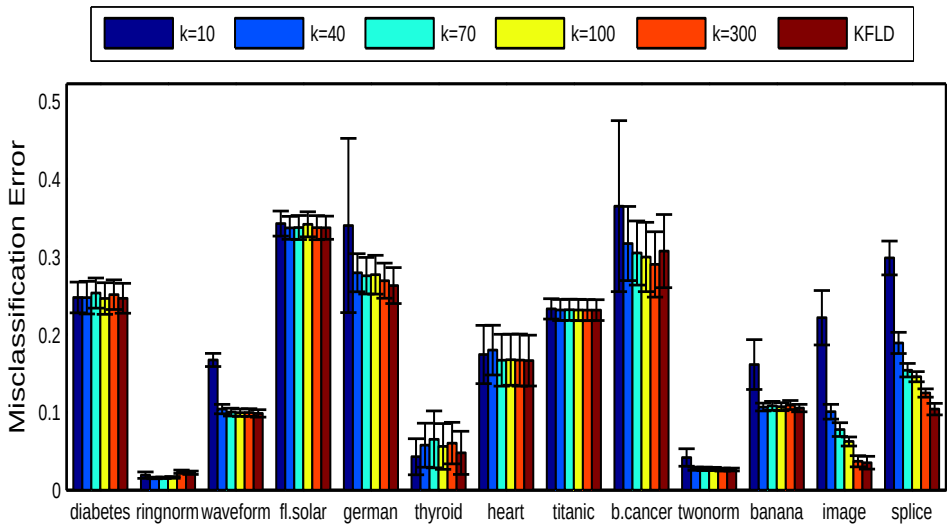

Fig. 3. Performance of randomly projected kernel-FLD classifiers on 13 UCI data sets.

with probability $1-\delta(\epsilon)$, where $\delta(\epsilon) = \exp\left(-\frac{Tr(\Sigma)}{2\lambda_{\max}(\Sigma)}(\sqrt{1+\epsilon}-1)^2\right)$. This can be proved with elementary techniques using the Laplace transform and the moment-generating function of a central χ^2 in ℓ_2 [14]; it also follows as a special case from the first part of Lemma 1 in [7] (where it was used for a different purpose).

Step 2 [Covering] Let $\mathcal{N}$ be an η-net over $\mathcal{S}^{k-1}$ with $\eta \in (0,1)$. Define

$$t := \frac{4k \cdot \lambda_{\max}(\Sigma)}{Tr(\Sigma)} \log(1+2/\eta) \tag{18}$$

Then, with probability $1-\exp(-k\log(1+2/\eta))$, we have uniformly $\forall \mathbf{w} \in \mathcal{N}$ that:

$$\frac{\|\Sigma^{1/2} R^T \mathbf{w}\|^2}{Tr(\Sigma)} \leqslant (1+\sqrt{t})^2 \tag{19}$$

Proof of step 2. The size of an η-net is bounded as $|\mathcal{N}| \leqslant (1+2/\eta)^k$ [18]. Applying eq.(17) from Step 1, and taking union bound over the points in $\mathcal{N}$ we have with probability $1-(1+2/\eta)^k\delta(\epsilon)$ that, $\forall \epsilon > 0$,

$$\frac{\|\Sigma^{1/2} R^T \mathbf{w}\|^2}{Tr(\Sigma)} \leqslant 1+\epsilon \tag{20}$$

We can make this probability large by an appropriate choice of ϵ. In particular, imposing $(1+2/\eta)^k\delta(\epsilon) = \delta^{1/2}(\epsilon)$, i.e.

$$(1+2/\eta)^k \exp\left(-\frac{Tr(\Sigma)}{2\lambda_{\max}(\Sigma)}(\sqrt{1+\epsilon}-1)^2\right) = \exp\left(-\frac{Tr(\Sigma)}{4\lambda_{\max}(\Sigma)}(\sqrt{1+\epsilon}-1)^2\right)$$

and solving this for ϵ gives:

$$1 + \epsilon = (1 + \sqrt{t})^2 \tag{21}$$

where t has been defined in eq.(18).

Finally, replacing this into eq.(20) and in $\delta(\epsilon)$ yields the statement of eq.(19) with probability $1 - \delta^{1/2}(\epsilon) = 1 - \exp(k \log(1 + 2/\eta))$ as required. □

Step 3 [Approximation] Let r be as in Step 2, and assume $t \in (0, 1)$. Then, uniformly over $\forall \mathbf{w} \in \mathcal{S}^{k-1}$, we have:

$$\frac{\mathrm{s}_{\max}(\Sigma^{1/2} R^T)}{\sqrt{Tr(\Sigma)}} \leqslant \frac{1}{1 - \eta}(1 + \sqrt{t}) \tag{22}$$

with probability $1 - \exp(-k \log(1 + 2/\eta))$.

Proof of step 3. Let $\mathbf{v} \in \mathcal{N}$ s.t. $\|\mathbf{w} - \mathbf{v}\| \leq \eta$. We have:

$$\frac{\|\Sigma^{1/2} R^T \mathbf{w}\|}{\sqrt{Tr(\Sigma)}} - 1 = \frac{\|\Sigma^{1/2} R^T \mathbf{w}\| - \|\Sigma^{1/2} R^T \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} + \frac{\|\Sigma^{1/2} R^T \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} - 1 \tag{23}$$

$$\leqslant \Big| \frac{\|\Sigma^{1/2} R^T \mathbf{w} - \Sigma^{1/2} R^T \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} \Big| + \frac{\|\Sigma^{1/2} R^T \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} - 1 \tag{24}$$

$$\leqslant \frac{\|\Sigma^{1/2} R^T\| \|\mathbf{w} - \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} + \frac{\|\Sigma^{1/2} R^T \mathbf{v}\|}{\sqrt{Tr(\Sigma)}} - 1 \tag{25}$$

$$\leqslant \frac{\|\Sigma^{1/2} R^T\|}{\sqrt{Tr(\Sigma)}} \eta + \sqrt{t} \tag{26}$$

where eq. (24) follows from the reverse triangle inequality, eq.(25) uses Cauchy-Schwartz, and eq.(26) follows by applying eq.(20) of Step 2 to the second term in eq.(25).

Note that $\|\Sigma^{1/2} R^T\|$ is the largest singular value of $\Sigma^{1/2} R^T$, and will be referred to as $\mathrm{s}_{\max}(\Sigma^{1/2} R^T)$.

Since eq.(26) holds uniformly $\forall \mathbf{w} \in \mathcal{S}^{k-1}$, it also holds for $\mathbf{w} := \arg\max_{\mathbf{w} \in \mathcal{S}^{k-1}} \|\Sigma^{1/2} R^T u\|$, i.e. the $\mathbf{w}$ for which $\|\Sigma^{1/2} R^T u\|$ achieves $\mathrm{s}_{\max}(\Sigma^{1/2} R^T)$. Using this, the r.h.s. inequality implies that:

$$\frac{\mathrm{s}_{\max}(\Sigma^{1/2} R^T)}{\sqrt{Tr(\Sigma)}} - 1 \leqslant \frac{\mathrm{s}_{\max}(\Sigma^{1/2} R^T)}{\sqrt{Tr(\Sigma)}} \eta + \sqrt{t} \tag{27}$$

hence

$$\frac{\mathrm{s}_{\max}(\Sigma^{1/2} R^T)}{\sqrt{Tr(\Sigma)}} \leqslant \frac{1}{1 - \eta}(1 + \sqrt{t}) \tag{28}$$

Rearranging, gives the statement of the theorem. □

Corollary 2. *With the notations and assumptions of Theorem 3, denote the* effective rank *of Σ by $r(\Sigma) := \frac{Tr(\Sigma)}{\lambda_{\max}(\Sigma)}$. Assume that $\frac{k}{r(\Sigma)}$ is bounded above*

by some positive constant $C > 0$. *Then,* $\forall \eta \in \left(0, \frac{\sqrt{k/r(\Sigma)}}{1+2\sqrt{\log 5}\cdot\sqrt{C}}\right]$, *we have with probability at least* $1 - \exp(-k\log(1+2/\eta))$*:*

$$\lambda_{\max}(R\Sigma R^T) \leqslant Tr(\Sigma)\left(1 + 4\sqrt{C\log(1+2/\eta)}\right)^2$$

Proof of Corollary 2. First, we apply Theorem 3 to $\mathrm{s}_{\max}(\Sigma^{1/2}R^T)$ with the choice $\eta = 1/2$:

$$\begin{aligned}\mathrm{s}_{\max}(\Sigma^{1/2}R^T) = \sqrt{\lambda_{\max}(R\Sigma R^T)} &\leqslant 2\sqrt{Tr(\Sigma)}\left(1 + 2\sqrt{\frac{k\cdot\lambda_{\max}(\Sigma)}{Tr(\Sigma)}\log 5}\right)\\ &\leqslant 2\sqrt{Tr(\Sigma)}\left(1 + 2\sqrt{C\log 5}\right) \qquad (29)\end{aligned}$$

Replacing this into eq. (26) we get:

$$\begin{aligned}\frac{\|\Sigma^{1/2}R^T\mathbf{w}\|}{\sqrt{Tr(\Sigma)}} - 1 \leqslant \frac{\|\Sigma^{1/2}R^T\|}{\sqrt{Tr(\Sigma)}}\eta + \sqrt{t} &\leqslant 2\left(1 + 2\sqrt{C\log 5}\right)\eta + \sqrt{t}\\ &\leqslant 2\left(1 + 2\sqrt{C\log 5}\right)\eta + 2\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)} \qquad (30)\end{aligned}$$

where in the last line we used the definition of t given in eq.(18).

Now, choose $0 < \eta \leqslant \frac{\sqrt{k/r(\Sigma)}}{1+2\sqrt{C\log 5}}$. This choice is valid, since it satisfies that $\frac{\sqrt{k/r(\Sigma)}}{1+2\sqrt{C\log 5}} \leqslant 1$ due to our precondition that $\frac{k}{r(\Sigma)} \leqslant C$.

With this choice, then the first term on the r.h.s. of eq.(30) becomes bounded as:

$$2\left(1 + 2\sqrt{C\log 5}\right)\eta \leqslant 2\sqrt{\frac{k}{r(\Sigma)}} \qquad (31)$$

This is smaller than the second term, $2\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)}$, since $\eta \leqslant 1$ (and so $\log(1+2/\eta) \geq \log 3 \geq 1$). Therefore in eq.(30) the second term dominates, and hence we can bound eq. (30) further by:

$$2\sqrt{\frac{k}{r(\Sigma)}} + 2\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)} \leqslant 4\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)} \qquad (32)$$

Summing up, we have uniformly $\forall \mathbf{u} \in \mathcal{N}$ that:

$$\frac{\|\Sigma^{1/2}R^T\mathbf{w}\|}{\sqrt{Tr(\Sigma)}} - 1 \leqslant 4\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)} \qquad (33)$$

It follows that:

$$\lambda_{\max}(RSR^T) \leqslant Tr(\Sigma)\left(1 + 4\sqrt{\frac{k}{r(\Sigma)}\log(1+2/\eta)}\right)^2 \qquad (34)$$

and using that $k \leqslant C \cdot r(\Sigma)$ concludes the proof. □

Proof of Theorem 2 We bound the error in the k-dimensional projection space, using Rayleigh quotient:

$$\hat{\Pr}_{(x_q,y_q)}[\hat{h}^R(Rx_q) \neq y_q] = \Phi\left(-\frac{1}{2}\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T [R\Sigma R^T]^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}\right)$$
$$\leqslant \Phi\left(-\frac{1}{2}\frac{\|R(\hat{\mu}_1 - \hat{\mu}_0)\|}{\sqrt{\lambda_{\max}(R\Sigma R^T)}}\right)$$

where we used that $\pi_0 + \pi_1 = 1$.

Now, applying Corollary 2 to the denominator, and applying the Hilbert-space version of Johnson-Lindenstrauss lemma [2] to the norm in the numerator completes the proof of claim a).

Finally, b) is obtained simply by applying union bound over the $m-1$ different ways that misclassification can occur, and the $m(m-1)/2$ distances between the m class centres. □

4 Conclusions

We have shown that Compressive FLD exhibits a dimension-adaptive property with respect to the random projection. We restricted ourselves to the analysis of the main term of the error in order focus on this property and we have shown that if the data density does not fill the ambient space then the error of compressive FLD can be bounded independently of the ambient dimension, with an expression that depends on a notion of 'intrinsic dimension' instead. In the case of data that lives in a linear subspace the intrinsic dimension is the dimension of that subspace. More generally, in the case of data whose class-conditional density has a trace-class covariance operator, the placeholder of the intrinsic dimension in our bound is the effective rank of the class covariance.

Due to the nice properties of random projections, and to many recent advances in this area, future work is aimed to derive learning guarantees that depend on some notions of complexity of the data geometry so that structural regularities that make learning easier should be reflected in better learning guarantees. As a by-product, learning in the randomly projected data space when the data density has regularities also leads to more efficient algorithms since the smaller the projected dimension is allowed to be the less computation time will be required.

References

1. R.I. Arriaga, S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. Proceedings of the 40-th Annual Symposium on Foundations of Computer Science (FOCS), 1999, pp. 616-623.
2. G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. IEEE Transactions on Information Theory, vol. 54, pp. 781-790, 2008.

3. S. Dasgupta. Learning mixtures of Gaussians. Proceedings of the 40-th Annual Symposium on Foundations of Computer Science (FOCS),1999, pp. 634-644.
4. S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Random Structures and Algorithms 22, pp. 60-65, 2002.
5. R.J. Durrant, A. Kabán. Compressed Fisher linear discriminant analysis: Classification of randomly projected data. Proceedings of the 16-th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2010.
6. R.J. Durrant and A. Kabán. A tight bound on the performance of Fisher's linear discriminant in randomly projected data spaces. Pattern Recognition Letters Volume 33, Issue 7, 1 May 2012, pp. 911-919, Special Issue on Awards from ICPR 2010.
7. R.J. Durrant, A. Kabán. Error bounds for kernel Fisher linear discriminant in Gaussian Hilbert space. 15-th International Conference on Artificial Intelligence and Statistics (AiStats), JMLR W&CP 22: 337-345, 2012.
8. R.J. Durrant, A. Kabán. Sharp Generalization Error Bounds for Randomly-projected Classifiers. 30th International Conference on Machine Learning (ICML 2013), JMLR W&CP 28 (3): 693-701, 2013.
9. A. Farahmand, Cs. Szepesvári, and J.-Y. Audibert. Manifold-adaptive dimension estimation, Proceedings of the 24th Annual International Conference on Machine Learning (ICML), 2007, pp. 265–272.
10. N. Halko, P.G. Martisson, J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review, Vol. 53, No. 2, 2011, pp. 217-288.
11. R.A. Horn, C.R. Johnson. Matrix analysis, CUP, 1985.
12. N.L. Johnson, S. Kotz and N. Balakrishnan. Continuous univariate distributions, Vol. 1. Wiley, 2 edition, 1994.
13. S Krishnan, C Bhattacharyya, R Hariharan. A randomized algorithm for large scale support vector learning. Proceedings of the 21-st Annual Conference on Neural Information Processing Systems (NIPS), 2007.
14. S. Maniglia and A. Rhandi. Gaussian measures on separable Hilbert spaces and applications. Quaderni del Dipartimento di Matematica dell' Universit del Salento, 2004, pages 1-24.
15. G.J. McLachlan. Discriminant analysis and statistical pattern recognition. 1992. Wiley.
16. S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and KR Mullers. Fisher discriminant analysis with kernels. Proc. of the 1999 IEEE Signal Processing Society Workshop, pages 41-48. IEEE, 2002.
17. T. Sarlós. Improved approximation algorithms for large matrices via random projections. Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2006, pp. 143-152.
18. R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012