JMLR: Workshop and Conference Proceedings 45:17–32, 2015 ACML 2015

Non-asymptotic Analysis of Compressive Fisher Discriminants in terms of the Effective Dimension

Ata Kabán AXK@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham,
Edgbaston, B15 2TT, Birmingham, UK

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

We provide a non-asymptotic analysis of the generalisation error of compressive Fisher linear discriminant (FLD) classification that is dimension free under mild assumptions. Our analysis includes the effects that random projection has on classification performance under covariance model misspecification, as well as various good and bad effects of random projections that contribute to the overall performance of compressive FLD. We also give an asymptotic bound as a corollary of our finite sample result. An important ingredient of our analysis is to develop new dimension-free bounds on the largest and smallest eigenvalue of the compressive covariance, which may be of independent interest.

Keywords: bounds on extreme eigenvalues, random projection, effective dimension, Fisher discriminant classification

1. Introduction

We consider the fundamental problem of 2-class classification, given a training set $\mathcal{T}_N = \{(x_n, y_n)\}_{n=1}^N$ sampled i.i.d. from a distribution $\mathcal{D}$ over the domain $\mathcal{X} \times \mathcal{Y}$. The input domain $\mathcal{X}$ can be taken as $\mathbb{R}^d$ with d arbitrarily large (more generally a separable Gaussian Hilbert space can be taken), and $\mathcal{Y} = \{0, 1\}$. For a given class of functions, $\mathcal{F}$, the goal 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 $\mathcal{L}$. We will use the $(0, 1)$-loss, which is the loss of interest in 2-class classification, so we can write the generalisation error of a classifier h as:

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

where (x, y) is a query point with unknown label y.

The first classification study is due to Fisher (1936). The method known under the name of Fisher Linear Discriminant (FLD) is still a widely used successful approach. Its simplicity made numerous analytical studies feasible, including very recent ones (McLachlan , 1992; Bian & Tao , 2014; Pourhabib et al. , 2015). There are also extensions of FLD to functional data spaces (James & Hastie, 2001; Shin , 2008).

We are interested in settings where the original data dimension d is arbitrarily large. Many machine learning methods are known to scale poorly when the dimension of the data space grows, and FLD is no exception. An interesting problem raised e.g. in Farahmand et al. (2007) is to devise algorithms whose performance scales with the hidden intrinsic dimension rather than the observed ambient dimension of the data.

Dimensionality reduction attempts to get around the problems of high dimension. Random projection is a universal method to do this, in the sense that it is oblivious to the data and has nice theoretical guarantees. Early work on compressive FLD – that is learning an FLD from random-projected data – only considered the conditional error when the training set is fixed, and has shown that in general the bounds, as well as the empirical performance, get worse with increasing d (Kabán & Durrant , 2013). Furthermore, in order to control the error of the compressive classifier, the compressive dimension is only required to be of the order logarithmic in the number of classes (Durrant & Kabán , 2012). In this paper we will look in more depth at the generalisation error of compressive FLD under such drastic dimensionality reduction, including the case when the shared covariance model of FLD may be misspecified.

Let R be a $k \times d$ matrix with entries drawn i.i.d. from a standard Gaussian. The projected training set that we work with has the following form: $\mathcal{T}_N^R = \{(Rx_i, y_i) : (x_i, y_i) \sim \mathcal{D}\}$. We seek to bound the probability that a query point $(Rx, y) \sim \mathcal{D}$ is misclassified by the compressive FLD. To this end, we will develop and apply new bounds on the extreme eigenvalues of the projected covariance that are dimension-free and depend only on the *effective dimension*.

Definition 1 *(Vershynin , 2011). Let Σ be a trace class covariance matrix in a separable Hilbert space, i.e. $Tr(\Sigma) < \infty$, and denote by $\lambda_{max}(\Sigma)$ its largest eigenvalue. The effective rank of Σ is defined as $r(\Sigma) = \frac{Tr(\Sigma)}{\lambda_{max}(\Sigma)}$.*

The remainder of the paper is structured in two main sections. The next section develops some tools that we will need. Section 3 then delves into a detailed and more complete analysis of the generalisation of compressive Fisher discriminant classification than previous analyses have done.

2. New Dimension-free Bounds on the Extreme Eigenvalues of a Wishart Matrix

2.1. Background

It was noted in Dasgupta (1999) that covariances become more spherical after orthogonal projection to a randomly oriented linear subspace. This fact is also quite intuitive to see: Let Σ be a covariance matrix in $\mathbb{R}^d$ with $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ denoting its largest and smallest eigenvalues respectively, and let R_o be a $k \times d$ random matrix with orthonormal rows, $k < d$. Then the Poincaré inequality (Horn & Johnson, 1985) says that $\lambda_{\max}(R_o \Sigma R_o^T) \leq \lambda_{\max}(\Sigma)$ and $\lambda_{\min}(R_o \Sigma R_o^T) \geq \lambda_{\min}(\Sigma)$.

It is often more convenient to use random matrices with i.i.d. entries in place of R_o; let R be a $k \times d$ random matrix with i.i.d. Gaussian or sub-Gaussian entries, $k < d$. When d is large enough, due to measure concentration, such a matrix R has almost orthogonal rows of almost the same length. For instance, if the entries are drawn i.i.d. from $\mathcal{N}(0, 1/d)$ then all rows of R have Euclidean norm close to 1. Moreover such random matrix will behave almost like the orthonormal R_o.

Then we can write

$$\lambda_{\max}(R\Sigma R^T) \leq \lambda_{\max}(\Sigma)\lambda_{\max}(RR^T), \text{ and } \lambda_{\min}(R\Sigma R^T) \geq \lambda_{\min}(\Sigma)\lambda_{\min}(RR^T), \quad (1)$$

and use the known high probability bounds on the largest and smallest eigenvalues of RR^T. The tightest bounds are known for R having i.i.d. Gaussian entries, and these match the Bai-Yin law at the limit when $d, k \to \infty$ and $k/d \to c \in [0, 1]$. The non-asymptotic bounds by Davidson & Szarek (2001) (Theorem II.13.) have been obtained using comparison inequalities for the suprema of Gaussian processes, namely the Slepian and Gordon inequalities. These results, given also in eq. (2.3) of Vershynin (2011), are the following:

Lemma 2 *(Davidson & Szarek , 2001) Let R be a $k \times d$ matrix with entries sampled i.i.d from $\mathcal{N}(0, 1)$. Then for all $\epsilon > 0$ with probability at least $1 - 2\exp(-\epsilon^2/2)$ we have:*

$$(\sqrt{d} - \sqrt{k} - \epsilon)_+^2 \leqslant \lambda_{min}(RR^T) \leqslant \lambda_{max}(RR^T) \leqslant (\sqrt{d} + \sqrt{k} + \epsilon)^2. \tag{2}$$

where the lower estimate requires that $k < d$.

Now, using these we can bound the extreme eigenvalues of $R\Sigma R^T$, where R is a $k \times d$ random matrix with $k < d$ and having i.i.d. entries from $\mathcal{N}(0, 1)$, simply as:

$$d{\cdot}\lambda_{\min}(\Sigma)(1-\sqrt{k/d}-\epsilon/\sqrt{d})_+^2 \leqslant \lambda_{\min}(R\Sigma R^T) \leqslant \lambda_{\max}(R\Sigma R^T) \leqslant d{\cdot}\lambda_{\max}(\Sigma)(1+\sqrt{k/d}+\epsilon/\sqrt{d})^2 \tag{3}$$

w.p. $1 - 2\exp(-\epsilon^2/2)$, where $(\cdot)_+ = \sup(\cdot, 0)$.

The problem occurs in settings when we would like to let d grow without bounds. On the surface, in eq. (3) the factor of d can be eliminated by choosing the entries of R to have variance $1/d$ instead of variance 1. However, if we do this then in return we will have $\mathrm{E}[\|Rx\|^2] = \mathrm{E}[\|R_o x\|^2] = \frac{k}{d}\|x\|^2$ for any $x \in \mathbb{R}^d$, so the dependence on d comes back again.

In other words, we cannot let d grow without bounds without either blowing up $\lambda_{\max}(R\Sigma R^T)$ or getting all distances shrink to zero after projection. These problems arise for example when trying to quantify the preservation of a Mahalanobis norm that is estimated from randomly projected data. In addition, if d grows unbounded and the infinite sequence of the eigenvalues of Σ converges to 0 while all of its finite eigenvalues are nonzero, then the lower bound on $R\Sigma R^T$ in eq. (3) cannot be used since there is no $\lambda_{\min}(\Sigma)$. This is the typical case for trace class covariances, which we will consider in this paper.

In the next subsection we get around both of these problems by deriving dimension-free bounds on the largest and smallest eigenvalues of $R\Sigma R^T$. To this end we will use the same tools as Davidson & Szarek (2001), i.e. the Slepian and Gordon inequalities, but we extend their proof to account for $\Sigma \neq I_d$. As we shall see, this allows us to make the bound independent of d. We state our bounds in R^d although this can be essentially replaced with a separable Hilbert space which can be identified with ℓ_2, equipped with Gaussian probability measure over Borel sets similar to (Biau et al., 2008). So d is allowed to be infinite as long as we require that Σ is a trace class covariance operator (i.e. $\mathrm{Tr}(\Sigma) < \infty$).

Before proceeding, the Slepian and Gordon inequalities that we will use (Ledoux & Talagrand (1991): pp.76-77; Davidson & Szarek (2001): Lemma II.9.) are given below.

Lemma 3 (Comparison inequalities) . *Consider two Gaussian processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ indexed by a bounded set T, with $E[X_t] = E[Y_t]$.*
(a) Slepian inequality*. If the expected increments satisfy:*

$$E[(X_t - X_{t'})^2] \leq E[(Y_t - Y_{t'})^2], \forall t, t' \in T, \tag{4}$$

then, $E[\sup_{t\in T} X_t] \leq E[\sup_{t\in T} Y_t]$.

(b) Gordon inequality. *If* $T = \cup_{s\in S} T_s$, *and the expected increments satisfy:*

$$E[(X_t - X_{t'})^2] \leq E[(Y_t - Y_{t'})^2], \forall t \in T_s, t' \in T_{s'}, s \neq s'; \text{ and} \quad (5)$$

$$E[(X_t - X_{t'})^2] \geq E[(Y_t - Y_{t'})^2], \forall t, t' \in T_s \text{ for some } s, \quad (6)$$

then, $E[\sup_{s\in S}\inf_{t\in T_s} X_{s,t}] \leq E[\sup_{s\in S}\inf_{t\in T_s} Y_{s,t}]$.

2.2. New bounds

Lemma 4 (Dimension-free bounds on the extreme eigenvalues of $R\Sigma R^T$) *Let* Σ *be a covariance matrix in* $\mathbb{R}^d$, *and we denote by* $\lambda_{max}(\cdot)$ *and* $\lambda_{min}(\cdot)$ *its largest and smallest eigenvalues. Let* R *be a* $k \times d$ *random matrix with i.i.d. standard Gaussian entries. For any* $\epsilon > 0$ *we have w.p. at least* $1 - \exp(-\epsilon^2/2)$*:*

$$\lambda_{max}(R\Sigma R^T) \leqslant \left(\sqrt{Tr(\Sigma)} + \sqrt{k \cdot \lambda_{max}(\Sigma)} + \epsilon\right)^2. \quad (7)$$

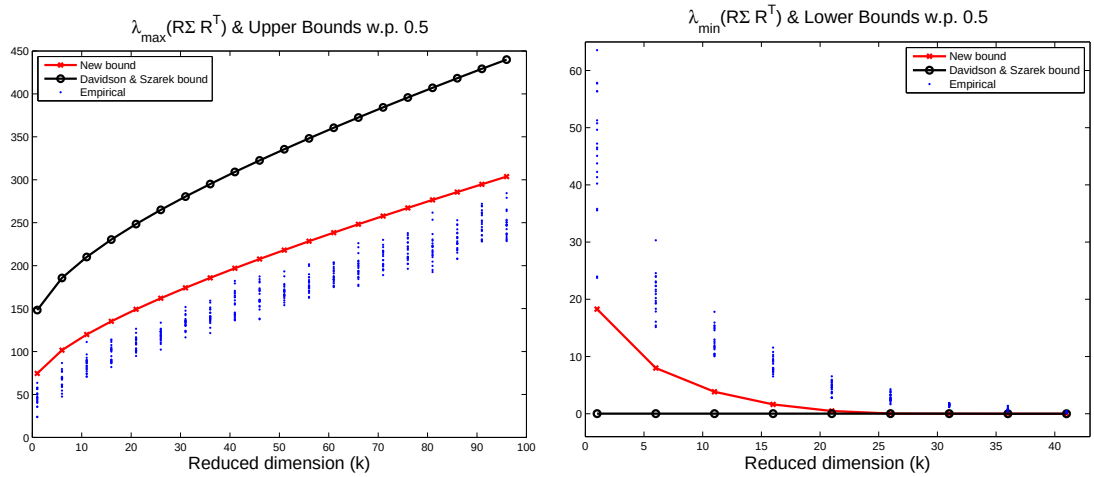
If $k < \lfloor \frac{Tr(\Sigma)}{\lambda_{max}(\Sigma)} \rfloor$ *then for any* $\epsilon \in (0, 1)$ *we have with probability at least* $1 - \exp(-\epsilon^2/2)$*:*

$$\lambda_{min}(R\Sigma R^T) \geqslant \left(\sqrt{Tr(\Sigma)} - \sqrt{k \cdot \lambda_{max}(\Sigma)} - \epsilon\right)_+^2. \quad (8)$$

Remark 5 *Let us observe that in the case of finite* d *and* $\Sigma = I_d$, *our bounds recover exactly the upper and lower estimates of Davidson & Szarek (2001), of which the upper bound on the largest eigenvalue is known to be sharp. In all other cases our bound in eq. (7) is tighter than that obtained in eq. (3) by the application of the bound of Davidson & Szarek (2001), because the effective dimension* $Tr(\Sigma)/\lambda_{max}(\Sigma)$ *is always no larger than the ambient dimension* d. *Furthermore, there are cases when the lower estimate in eq. (8) is also tighter than the corresponding bound in eq. (3).*

Example 1 *Figure 1 shows an illustration of these new bounds against the empirical distributions of the extreme eigenvalue estimates of* $R\Sigma R^T$, *in comparison with the simpler bounds in eq. (3). The* $d \times d$ *covariance matrix* Σ *has its first 40 eigenvalues equal to 1 and the remaining eigenvalues decay as the sequence* $(1/i^2)_{i=1,\ldots,d-40}$. *In the numerical simulation,* $d = 100$, *but by construction the trace of* Σ *is upper bounded by* $40 + \pi^2/6$ *for any arbitrarily large* d. *The improved tightness is most apparent on these figures.*

We should note though that the proof relies on the use of comparison inequalities for the suprema of Gaussian processes, and this implies that the bounds in Lemma 4, just like those of Davidson & Szarek (2001), are specific to Gaussian R. It is not clear whether similarly tight bounds could be obtained for subgaussian R. The largest singular value of a product of deterministic and random matrix was studied under much more general non-Gaussian R in Vershynin (2011) by other techniques yielding bounds of similar form but with worse constants and/or an additional logarithmic term. Those bounds are also independent of d



(a) (b)

Figure 1: Illustration of the bounds given in Lemma 4. The $d \times d$ covariance matrix Σ has its first 40 eigenvalues equal to 1 and the remaining eigenvalues decay as the sequence $(1/i^2)_{i=1,\ldots,d-40}$. In this simulation, $d = 100$. The bounds from eq.(3) that use Davidson & Szarek (2001) are shown for comparison, along with the empirical distributions of the extreme eigenvalue estimates of $R\Sigma R^T$.

and depend on the rank of Σ instead.

Proof (Sketch) of Lemma 4, eq. (7).
Without loss of generality we can identify Σ with the diagonal matrix of its eigenvalues Λ by absorbing its eigenvectors into the Gaussian sequences $(R_{ji})_{i\geqslant 1}, j = 1, \ldots k$. Then write the largest singular value of $\Lambda^{1/2}R^T$ in the following form:

$$\mathrm{s}_{\max}(\Lambda^{1/2}R^T) = \sup_{u\in\mathcal{S}^{k-1}, v\in\mathcal{S}^{d-1}} u^T R\Lambda^{1/2} v \tag{9}$$

where $\mathcal{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$.

Taking the index set $T = \mathcal{S}^{k-1}\times\mathcal{S}^{d-1}$ with elements $t = (u, v) \in T$, we define a Gaussian process as the following:

$$X_{uv} = u^T R\Lambda^{1/2} v. \tag{10}$$

The supremum of X_{uv} is the singular value of our interest.

Now, the strategy is to compute the expected squared increments of X_{uv}, and upper-bound it with a quantity that coincides with the expected squared increments of another Gaussian process whose supremum is easy to compute.

It can be verified (details omitted) that the following definition fits the bill for the second Gaussian process:

$$Y_{uv} = \sqrt{\lambda_{\max}(\Sigma)}\cdot u^T g + v^T h \tag{11}$$

where $g \sim \mathcal{N}(0, I_k)$, and $h \sim \mathcal{N}(0, \Lambda)$ are independent of each other. Note that h is well defined even if d is allowed to be infinite because, Λ is trace-class. So by the Slepian

inequality, after some working, we can get the following upper bound on the expected largest singular value of $\Lambda^{1/2}R^T$:

$$\mathrm{E}[s_{\max}(\Lambda^{1/2}R^T)] \leq \sqrt{Tr(\Sigma)} + \sqrt{k \cdot \lambda_{\max}(\Sigma)}. \tag{12}$$

The final step is to bound $s_{\max}(\Lambda^{1/2}R^T)$ away from its expectation, using the fact that $s_{\max}(\cdot)$ is 1-Lipschitz w.r.t. the spectral norm (Vershynin (2011): Corollary 5.35). Hence, invoking the the concentration of Lipschitz functions in Gaussian space we get:

$$\mathrm{s}_{\max}(\Lambda^{1/2}R^T) \leqslant \sqrt{Tr(\Sigma)} + \sqrt{k \cdot \lambda_{\max}(\Sigma)} + \epsilon \tag{13}$$

with probability at least $1 - \exp(-\epsilon^2/2)$. This completes the proof of eq.(7) □

Proof (Sketch) of Lemma 4, eq.(8).

We denote by $\mathrm{s}_{\inf}$ the infimum of a sequence of singular values. We use infimum here to accommodate the fact that d, and hence the sequence of singular values of $\Lambda^{1/2}R^T$, is allowed to be infinite.

It is useful to write $\mathrm{s}_{\inf}$ of $\Lambda^{1/2}R^T$ in the following form:

$$\mathrm{s}_{\inf}(\Lambda^{1/2}R^T) = - \sup_{u \in \mathcal{S}^{k-1}} \inf_{v \in \mathcal{S}^{d-1}} u^T R \Lambda^{1/2} v. \tag{14}$$

Then take the index set $T = \mathcal{S}^{k-1} \times \mathcal{S}^{d-1}$ with elements denoted as $t = (u, v) \in T$ as before, and take $S = \mathcal{S}^{k-1}$ with elements u. For each $u \in S$ define $T_u = \{(u, v) : v \in \mathcal{S}^{d-1}\}$. It is easy to see that $T = \cup_{u \in S} T_u$.

With the same definitions of X_{uv} and Y_{uv} as given before in eqs (10) and (11), one can verify that the two conditions on the Gordon inequality hold. Thus, from the Gordon inequality we have that the negative of the smallest singular value of our interest, i.e. $\mathrm{E}[\sup_{u \in S} \inf_{t \in T_u} X_{uv}]$ is upper bounded by $\mathrm{E}[\sup_{u \in S} \inf_{t \in T_u} Y_{uv}]$, where Y_{uv} is, as in eq. (11) – which works out as:

$$\mathrm{E}[\sup_{u \in S} \inf_{t \in T_u} Y_{uv}] = \sqrt{\lambda_{\max}(\Sigma)} \cdot \mathrm{E}[\|g\|] - \mathrm{E}\left[\|h\|\right] \tag{15}$$

with $g \sim \mathcal{N}(0, I_k)$, and $h \sim \mathcal{N}(0, \Lambda)$ independent of each other.

Now it remains to compute (or upper bound) the expectations in eq.(15), and the negative of the r.h.s. of eq.(15) (or an upper bound on it) will give us the lower bound on $\mathrm{s}_{\inf}(\Lambda^{1/2}R^T)$ of our interest.

The first term is the expectation of a χ variable, that is,

$$\mathrm{E}_{g \sim \mathcal{N}(0, I_k)}[\|g\|] = \sqrt{2}\frac{\Gamma((k+1)/2)}{\Gamma(k/2)}. \tag{16}$$

The second term does not have a known form, and requires some work. Using recent results from Pinelis (2015a), and references therein, we can bound this from below with $\sqrt{\lambda_{\sup}(\Sigma)}$ times the expectation of a Nakagami distribution as the following:

$$\mathrm{E}[\|h\|] \geq \sqrt{\lambda_{\max}(\Sigma)}\sqrt{2}\frac{\Gamma((r(\Sigma)+1)/2)}{\Gamma(r(\Sigma)/2)} \tag{17}$$

where $r(\Sigma)$ is the effective dimension.

Now, since $k \leq r(\Sigma)$, we can use, as in Davidson & Szarek (2001): Thm 2.13, that the sequence $\left(\sqrt{n} - \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\right)_{n\geq 1}$ is decreasing. Therefore it follows that the r.h.s. of eq.(15) is bounded as:

$$\sqrt{\lambda_{\max}(\Sigma)}(\mathrm{E}[\|g\|] - \mathrm{E}[\|h\|]) \quad \leq \quad \sqrt{k \cdot \lambda_{\max}(\Sigma)} - \sqrt{\mathrm{Tr}(\Sigma)} \tag{18}$$

Plugging this back into eqs. (15) and (14) we get the statement in eq. (8).

The final step is again to note that $s_{\inf}(.)$ is 1-Lipschitz with respect to the spectral norm (Vershynin (2011) Corollary 5.35) and use Gaussian concentration of Lipschitz functions around their expectation. □

3. Application to bounding the error of compressive Fisher discriminant classification in terms of the *effective dimension*

In this section the class of functions $\mathcal{F}$ will consist of FLD classifiers. As elsewhere in the literature (Bian & Tao , 2014; Pourhabib et al. , 2015), we will assume that the two classes follow multivariate Gaussian distributions – however, we do not assume identical class covariances in the true data distribution. The only requirement we need is that the true class covariances have finite trace. We then model this data by an FLD model, that is, the model covariance that is shared between the two classes. Part of our analysis will quantify the effects of this covariance misspecification.

The class label of a query point is predicted according to the smallest Mahalanobis distance from the class centers. Denoting by $\hat{\Sigma}$ the maximum likelihood estimate of the pooled covariance and by $\hat{\mu}_0$ and $\hat{\mu}_1$ the maximum likelihood estimates of the class means on the original data, the decision function of compressive FLD, denoted as $\hat{h}^R$, for an input query point x is:

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

Our interest is the generalisation error (or misclassification error) of $\hat{h}^R$, which is defined as $\Pr_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N^R]$. We will prove the following result.

Theorem 6 *[Non-asymptotic generalisation error bound for compressive FLD] Let $(x, y) \sim \mathcal{D}$ be a query point with unknown label y and Gaussian class conditionals $x|y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$ with trace class covariances, i.e. $Tr(\Sigma_i) < \infty, \forall i = \{0, 1\}$. Let $\pi_i = Pr(y = i)$ be bounded away from both 0 and 1. Let R be a $k \times d$ random matrix with i.i.d. standard Gaussian entries. Then, $\forall \epsilon \in (0, 1)$, the generalisation error is bounded as the following:*

$$Pr_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N] \leqslant ...$$

$$\sum_{i=0}^{1} \pi_i \Phi\left(-\left[\left[\sqrt{k} - \epsilon\right]_+ \frac{\left[\sqrt{\|\mu_1 - \mu_0\|^2 + \frac{Tr(\beta_1\Sigma_0 + \beta_0\Sigma_1)}{N\alpha_0\alpha_1}} - \epsilon\sqrt{\frac{\lambda_{max}(\alpha_1\Sigma_0 + \alpha_0\Sigma_1)}{N\beta_0\beta_1}}\right]_+}{\sqrt{Tr(\Sigma_i)} + \sqrt{k\lambda_{max}(\Sigma_i)} + \epsilon} g(\tilde{\kappa}_i) - \frac{\sqrt{k} + \epsilon}{\sqrt{N\beta_i}}\right]_+\right) \tag{19}$$

w.p. $1-10\exp(-\epsilon^2/2)-2\exp(-\pi_0 N\epsilon^2/3)$, *where* Φ *is the standard Gaussian cumulative distribution function,* $\alpha_0=\pi_0(1+\epsilon)$, $\alpha_1=1-\pi_0(1-\epsilon)$, $\beta_0=\pi_0(1-\epsilon)$ *and* $\beta_1=1-\pi_0(1+\epsilon)$, *and* $g(\tilde{\kappa}_i)=\frac{\sqrt{\tilde{\kappa}_i}}{1+\tilde{\kappa}_i}$.

In the case when $\Sigma_0=\Sigma_1=\Sigma$, *then for* $k<(\sqrt{N-2}-\epsilon)^2$ *we have:*

$$\tilde{\kappa}_i=\left(\frac{\sqrt{N-2}+\sqrt{k}+\epsilon}{\sqrt{N-2}-\sqrt{k}-\epsilon}\right)^2. \tag{20}$$

In the case when $\Sigma_0\neq\Sigma_1$, *then*

$$\tilde{\kappa}_i=\frac{\left[(\sqrt{N-2}+\sqrt{k}+\epsilon)^2+(\sqrt{N\alpha_{\neg i}-1}+\sqrt{k}+\epsilon)^2\tilde{\lambda}_{max}(M_i)-(\sqrt{N\beta_{\neg i}-1}-\sqrt{k}-\epsilon)_+^2\right]_+}{\left[(\sqrt{N-2}-\sqrt{k}-\epsilon)_+^2+(\sqrt{N\beta_{\neg i}-1}-\sqrt{k}-\epsilon)_+^2\tilde{\lambda}_{min}(M_i)-(\sqrt{N\alpha_{\neg i}-1}+\sqrt{k}+\epsilon)^2\right]_+}, \tag{21}$$

provided that k *and* N *are such that this is finite. In the above,*

$$\tilde{\lambda}_{max}(M_i) = min\left\{\frac{(\sqrt{Tr(\Sigma_{\neg i})}+\sqrt{k\cdot\lambda_{max}(\Sigma_{\neg i})}+\epsilon)^2}{(\sqrt{Tr(\Sigma_i)}-\sqrt{k\cdot\lambda_{max}(\Sigma_i)}-\epsilon)_+^2},\lambda_{max}(\Sigma_i^+\Sigma_{\neg i})\right\} \tag{22}$$

$$\tilde{\lambda}_{min}(M_i) = max\left\{\frac{(\sqrt{Tr(\Sigma_i)}-\sqrt{k\cdot\lambda_{max}(\Sigma_i)}-\epsilon)_+^2}{(\sqrt{Tr(\Sigma_{\neg i})}+\sqrt{k\cdot\lambda_{max}(\Sigma_{\neg i})}+\epsilon)^2},\lambda_{min}(\Sigma_i^+\Sigma_{\neg i})\right\} \tag{23}$$

where $(\cdot)^+$ stands for any choice of generalised inverse.

Furthermore, by letting the sample size $N\to\infty$ we get the following asymptotic error bound.

Corollary 7 (Asymptotic generalisation error bound for compressive FLD) . *Under the conditions of Theorem 6, and the same* $g(\cdot)$, *we have:*

$$\limsup_{N\to\infty} Pr_{x,y}[\hat{h}^R(Rx)\neq y|\mathcal{T}_N]\leqslant\sum_{i=0}^{1}\pi_i\Phi\left(-\frac{(\sqrt{k}-\epsilon)\|\mu_1-\mu_0\|}{\sqrt{Tr(\Sigma_i)}+\sqrt{k\cdot\lambda_{max}(\Sigma_i)}+\epsilon}\cdot g\left(\frac{\tilde{\lambda}_{max}(M_i)}{\tilde{\lambda}_{min}(M_i)}\right)\right) \tag{24}$$

w.p. $1-4\exp(-\epsilon^2/2)$, *where* $\tilde{\lambda}_{max}(M_i)$ *and* $\tilde{\lambda}_{min}(M_i)$ *are as given in eqs. (22)-(23).*

Clearly, as expected, eq. (24) is an upper bound on the Bayes error since both the misspecification of a shared covariance and the random projection introduce biases. However the distortion is controlled, and the bound has the main characteristics of the FLD error. In particular, the crucial role of the distance between the class centers is most apparent; the $g()$ function encodes the price to pay for a misspecification of a shared covariance. If we divide through both the numerator and the denominator by $\lambda_{\max}(\Sigma_i)$ we can see that the denominator scales with the effective dimension $r(\Sigma_i)$, and the numerator scales with the distance between the class centers relative to $\lambda_{\max}(\Sigma_i)$. The condition number in the $g(\cdot)$ function in the asymptotic bound evaluates to $1/2$ if and only if the true class covariances are identical, and it is less than $1/2$ otherwise, so the error bound increases with the increase of this condition number.

The finite sample bound in Theorem 6 shares the same main characteristics as above, but of course it provides finer details since it holds true for any training set of size N. Its various components and their behaviours will be exemplified in numerical simulations shortly.

3.1. Proof (Sketch) of Theorem 6, and interpretation of the bounds

We start by writing the generalisation error conditional on both R and $\mathcal{T}_N$, and decomposing in a similar fashion as in (Durrant & Kabán , 2012). Define the following three terms:

$$\begin{aligned} A_i &:= \|(R\Sigma_i R^T)^{-\frac{1}{2}} R\,(\hat{\mu}_1 - \hat{\mu}_0)\,\| \\ B_i &:= \frac{\sqrt{\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})}}{1+\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-\frac{1}{2}})} \\ C_i &:= \|(R\Sigma_i R^T)^{-\frac{1}{2}} R(\mu_i - \hat{\mu}_i)\| \end{aligned}$$

Using these terms, we can decompose the generalisation as the following:

$$\begin{aligned} &\Pr_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N, R] = ... \\ &\sum_{i=0}^{1} \pi_i \Phi\left(-\frac{1}{2} \frac{(\hat{\mu}_{\neg i} - \hat{\mu}_i)^T R^T (R\hat{\Sigma}R^T)^{-1} R\,(\hat{\mu}_{\neg i} + \hat{\mu}_i - 2\mu_i)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}} \right) = ... \\ &\sum_{i=0}^{1} \pi_i \Phi\left(-\frac{1}{2} \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\,(\hat{\mu}_1 - \hat{\mu}_0) - 2\,(\mu_i - \hat{\mu}_i)^T R^T (R\hat{\Sigma}R^T)^{-1} R\,(\hat{\mu}_{\neg i} - \hat{\mu}_i)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T R^T (R\hat{\Sigma}R^T)^{-1} R\Sigma_i R^T (R\hat{\Sigma}R^T)^{-1} R(\hat{\mu}_1 - \hat{\mu}_0)}} \right) \leqslant \\ &\sum_{i=0}^{1} \pi_i \Phi\left(-[A_i B_i - C_i]\right) \end{aligned} \tag{25}$$

where eq. (25) follows from Kantorovich inequality (Horn & Johnson, 1985) applied to the first term of the numerator and Cauchy-Schwarz to the second.

In what follows, we will lower-bound A_i and B_i and upper-bound C_i.

Before proceeding, we give the following tail bound that we will make use of.

Lemma 8 (Dimension-free bound on the norm of a (sub)Gaussian vector) *Let $X \sim \mathcal{N}(\mu, \Sigma)$. For any $\epsilon \in (0,1)$,*

$$Pr\left\{ \left| \|X\| - \sqrt{Tr(\Sigma) + \|\mu\|^2} \right| \geq \epsilon \sqrt{\lambda_{max}(\Sigma)} \right\} \leq 2\exp\left\{ -\frac{\epsilon^2}{2} \right\} \tag{26}$$

This follows directly from Gaussian concentration of Lipschitz functions, since $\|\cdot\|$ is $\lambda_{\max}(\Sigma)$-Lipschitz. Interestingly, it also holds more generally for sub-Gaussian X, as a consequence of the Hanson-Wright inequality (Rudelson & Vershynin , 2013). An elementary proof can also be found in Lemma 1 of Durrant & Kabán (2012), which, after rearranging terms, turns out to be identical to Lemma 8 above.

3.2. Lower bound on A_i

The term A_i is the distance between the class center estimates in the Mahalanobis metric defined by the true covariance of class i. A larger value for this term implies better generalisation.

Using a combination of the Rayleigh quotient inequality, together with Lemma 4, and Lemma 8 applied twice, we arrive at the following lower bound, w.p. $1-\exp(-\epsilon_1^2/2)-\exp(-\epsilon_{2,i}^2/2)-\exp(-\epsilon_3^2/2)$:

$$A_i \geqslant \frac{[\sqrt{k}-\epsilon_1]_+}{\sqrt{Tr(\Sigma_i)}+\sqrt{k}\sqrt{\lambda_{\max}(\Sigma_i)}+\epsilon_{2,i}} \cdot \left[\sqrt{\|\mu_1-\mu_0\|^2+\mathrm{Tr}(\frac{\Sigma_0}{N_0}+\frac{\Sigma_1}{N_1})}-\epsilon_3\sqrt{\lambda_{\max}(\frac{\Sigma_0}{N_0}+\frac{\Sigma_1}{N_1})}\right]_+ \tag{27}$$

One can also verify that a similar strategy for bounding the corresponding $A_i^{\text{Dataspace FLD}} = \|\Sigma_i^{-1/2}(\hat{\mu}_1-\hat{\mu}_0)\|$ term of the dataspace FLD would yield exactly the expression in the bracket of second factor in eq. (27), w.p. $1-\exp(-\epsilon_3^2/2)$, and so the fraction in the first factor essentially encodes the impact that random projection has on this term. Unsurprisingly, this fraction decreases with k, deteriorating the error as k gets smaller. Hence, from the analysis of A_i terms we see that the data distributions that have large relative distance between their class centers are the ones that allow a more drastic random compression without causing too much unwanted deterioration of the classification accuracy. Feature space representations via the kernel trick may yield such data distributions.

Figure 2 illustrates the behaviour of an A_i term. For Σ_0 we used the same covariance matrix as in the earlier simulation, and Σ_1 was an arbitrary rotation of Σ_0. The true centeres μ_0, μ_1 were set to arbitrary locations at Euclidean distance 14.1 of each other. We see, especially after zooming in, that our lower bound on A_i increases with increasing k, in agreement with the behaviour observed in the empirical data distriburions. However the bound is much tighter for small values of k than it is for higher k values. The corresponding term $A_i^{\text{Dataspace FLD}}$ of the dataspace FLD is also plotted for reference.

3.3. Lower bound on B_i

Each term B_i is a decreasing function of the condition number of a matrix that encodes the mismatch between the true covariance of the i-th class and the pooled covariance estimate of the model – this includes contributions from both covariance misestimation from finite samples, and from misspecification of a shared covariance model. A higher value of this function is better, the highest achievable being 0.5. As we shall see, the random projection actually helps reduce this mismatch, and the more we compress the data the smaller the condition number, so the higher the value of the B_i term.

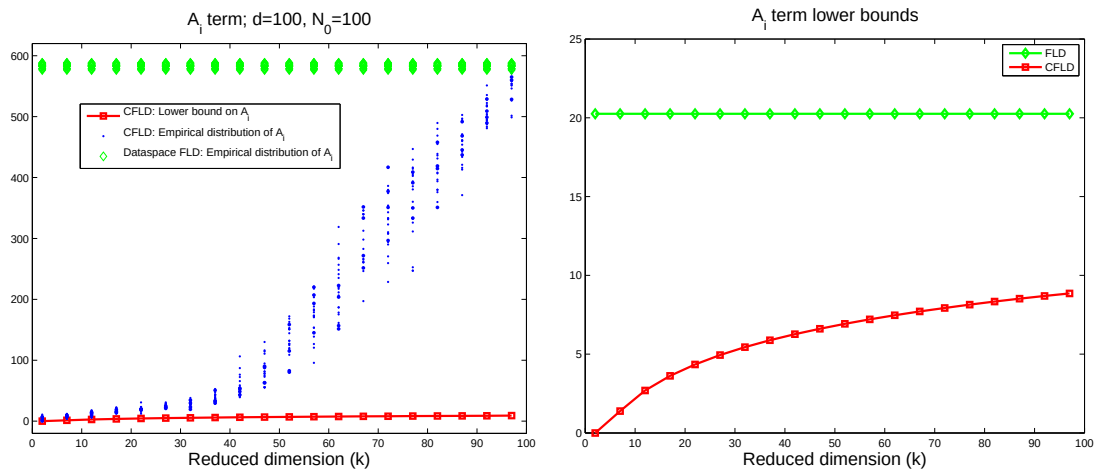
Recall, the true class covariances Σ_i and $\Sigma_{\neg i}$ are trace class. We can show w.p. $1-4\exp(-\epsilon_4^2/2)$ the following:

$$\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}}R\Sigma_i R^T(R\hat{\Sigma}R^T)^{-\frac{1}{2}}) \leq \ldots$$
$$\frac{\left[(\sqrt{N-2}+\sqrt{k}+\epsilon_4)^2+(\sqrt{N_{\neg i}-1}+\sqrt{k}+\epsilon_4)^2\tilde{\lambda}_{\max}(M_i)-(\sqrt{N_{\neg i}-1}-\sqrt{k}-\epsilon_4)_+^2\right]_+}{\left[(\sqrt{N-2}-\sqrt{k}-\epsilon_4)_+^2+(\sqrt{N_{\neg i}-1}-\sqrt{k}-\epsilon_4)_+^2\tilde{\lambda}_{\min}(M_i)-(\sqrt{N_{\neg i}-1}+\sqrt{k}+\epsilon_4)^2\right]_+} =: \tilde{\kappa}_i \tag{28}$$

where

$$M_i = (R\Sigma_i R^T)^{-1/2}R\Sigma_{\neg i}R^T(R\Sigma_i R^T)^{-1/2} \tag{29}$$

encodes the mismatch between the two true class-conditional covariances after random projection. The notations $\tilde{\lambda}_{\max}(M_i)$ and $\tilde{\lambda}_{\min}(M_i)$ represent upper and lower bounds on the extreme eigenvalues of M_i, given in subsection 3.3.1. The simple special $\Sigma_0 = \Sigma_1$ (i.e. the case of correct model specification) is given in subsection 3.3.2.



(a) (b)

Figure 2: Illustration of the lower bound on an A_i term. (a) The empirical distributions are superimposed as obtained from 20 independent realisations of the training set and the R. The corresponding empirical estimate for the A_i term for the dataspace FLD is also shown for comparison and reference. (b) The same lower bound on A_i zoomed in, along with a lower bound on $A_i^{\text{Dataspace FLD}}$ derived in the same way.

3.3.1. Bounding the extreme eigenvalues of M_i

The extreme eigenvalues of the matrix M_i represent the error contribution of FLD's simplifying model assumption of a shared class-covariance when the true class covariances are in fact different.

In the finite dimensional case the separation theorem for generalised eigenvalues (Scott & Styan , 1985) can be used, provided that Σ_0 and Σ_1 share the same column space:

$$\lambda_{\max}(M_i) \leq \lambda_{\max}(\Sigma_i^+ \Sigma_{\neg i}); \qquad \lambda_{\min}(M_i) \geq \lambda_{\min}(\Sigma_i^+ \Sigma_{\neg i}) \tag{30}$$

where $(\cdot)^+$ stands for any choice of generalised inverse.

Note that the inequalities in eq. (30) imply that this simplifying assumption is less damaging in the compressive space than in the original data space. In other words, random projection helps ameliorate the covariance misspecification.

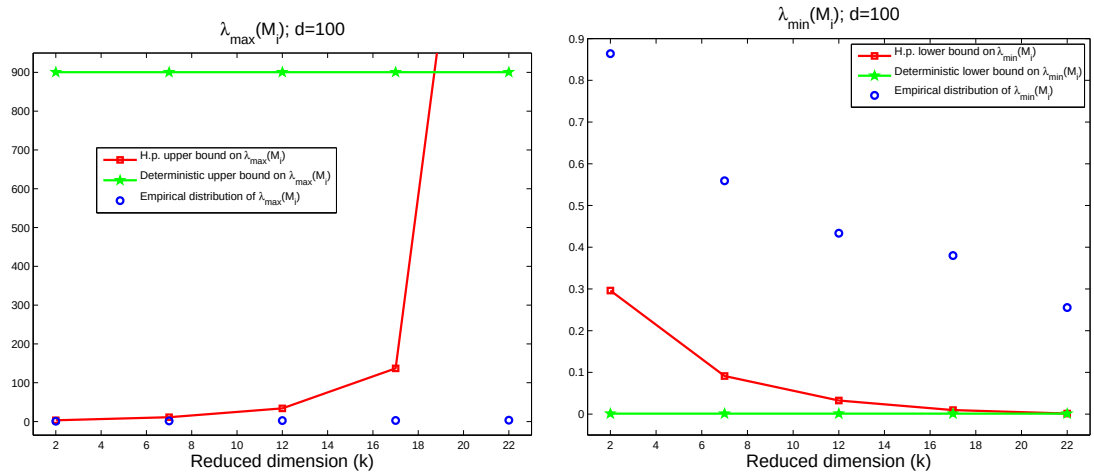
However, the deterministic inequalities in eq. (30) can sometimes be loose in high dimensions if $\Sigma_i^+ \Sigma_{\neg i}$ has large condition number. An alternative is to use the bounds developed in the previous section, which yield the following:

$$\lambda_{\max}(M_i) \quad \leq \quad \frac{\lambda_{\max}(R\Sigma_{\neg i}R^T)}{\lambda_{\min}(R\Sigma_i R^T)} \leq \frac{\left(\sqrt{Tr(\Sigma_{\neg i})} + \sqrt{k \cdot \lambda_{\max}(\Sigma_{\neg i})} + \epsilon\right)^2}{\left(\sqrt{Tr(\Sigma_i)} - \sqrt{k \cdot \lambda_{\max}(\Sigma_i)} - \epsilon\right)_+^2} \tag{31}$$

$$\lambda_{\min}(M_i) \quad \geq \quad \frac{\lambda_{\min}(R\Sigma_{\neg i}R^T)}{\lambda_{\max}(R\Sigma_i R^T)} \geq \frac{\left(\sqrt{Tr(\Sigma_{\neg i})} - \sqrt{k \cdot \lambda_{\max}(\Sigma_{\neg i})} - \epsilon\right)_+^2}{\left(\sqrt{Tr(\Sigma_i)} + \sqrt{k \cdot \lambda_{\max}(\Sigma_i)} + \epsilon\right)^2} \tag{32}$$

The latter bounds may be tighter in some cases, especially when k is small. In Figure 3 we illustrate this, where we used the same covariances as in the previous simulations, with the second covariance being a random rotation of the first. We then take the minimum and maximum between the corresponding $\lambda_{\max}(M_i)$ and $\lambda_{\min}(M_i)$ bounds respectively before replacing into eq.(28).

We see from the figures that random projection is beneficial for these extreme eigenvalue terms. This is because random projection diminishes the differences between the two class-conditional covariances and improves the condition number of M_i.



(a) (b)

Figure 3: Illustration of the bounds on the extreme eigenvalues of M_i. For the largest eigenvalue smaller values are better; for the smallest eigenvalue larger values are better. we see that the deterministic bound tends to be loose than the high probability bound when k is small, and tighter when k is large. We also see that random projection helps achieve better values for these terms.

3.3.2. The special case of $\Sigma_0 = \Sigma_1$

Returning to the task of lower bounding of B_i terms, this is much simpler when there is no model misspecification, so the class-conditional covariances are in fact identical. In this special case, $\Sigma_i = \Sigma_{\neg i} = \Sigma$, we can get a neater bound on the condition number in B_i as the following:

$$\kappa((R\hat{\Sigma}R^T)^{-\frac{1}{2}}R\Sigma R^T(R\hat{\Sigma}R^T)^{-\frac{1}{2}}) \quad \leq \quad \left(\frac{\sqrt{N-2}+\sqrt{k}+\epsilon_4}{\sqrt{N-2}-\sqrt{k}-\epsilon_4}\right)^2 \tag{33}$$

w.p. $1 - 2\exp(-\epsilon_4^2/2)$ w.r.t. the random draws of $\mathcal{T}_N$.

This completes the lower bounds on B_i. Figure 4 puts all these pieces together and shows the empirical behaviour of a B_i term along with our bounds. In this simulation the true class-conditional covariances are in fact identical, and we plot both our specialised bound for this situation, i.e. eq. (33) plugged into the function $g(\cdot)$, and the bound that we derived for the more general misspecified situation, i.e. eq. (28) plugged into $g(\cdot)$. As expected, the specialised bound is tighter – in fact, as we can see, it tightly follows the empirical behaviour. The more general bound also follows the main trend of the empirical behaviour, albeit it is understandably slightly looser. The empirical behaviour of the corresponding dataspace FLD $B_i^{\text{Dataspace FLD}}$ term is also shown, and we can see that the compressed version achieves higher (better) values for this term. This is because the compression acts as a regularisation by which both covariance misestimation and covariance misspecification effects are ameliorated.

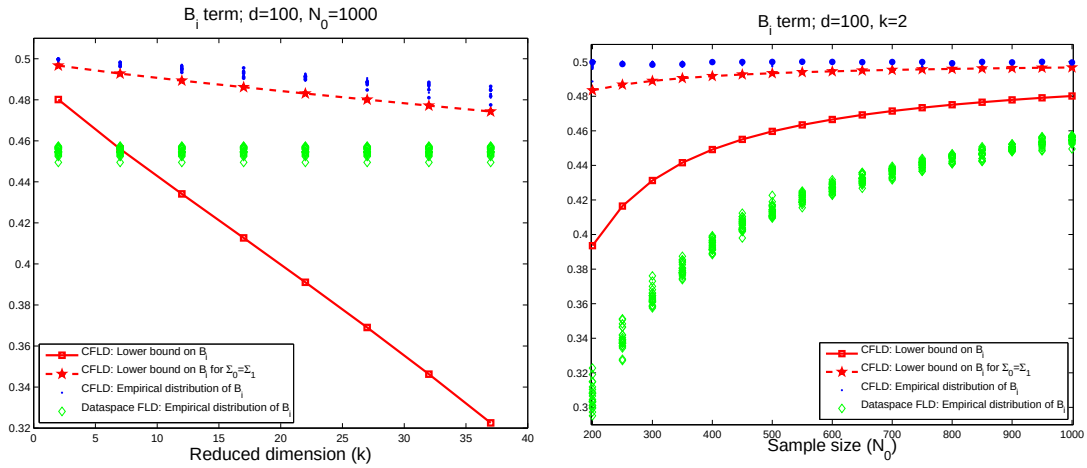


Figure 4: Illustration of the lower bounds on a B_i term. Higher values are better. We see that random projection has a very beneficial effect on this term. We also see that both of our lower bounds reflect the empirical behaviour faithfully, and as expected, the more specialised bound for equal covariances is tighter.

3.4. Bounding C_i

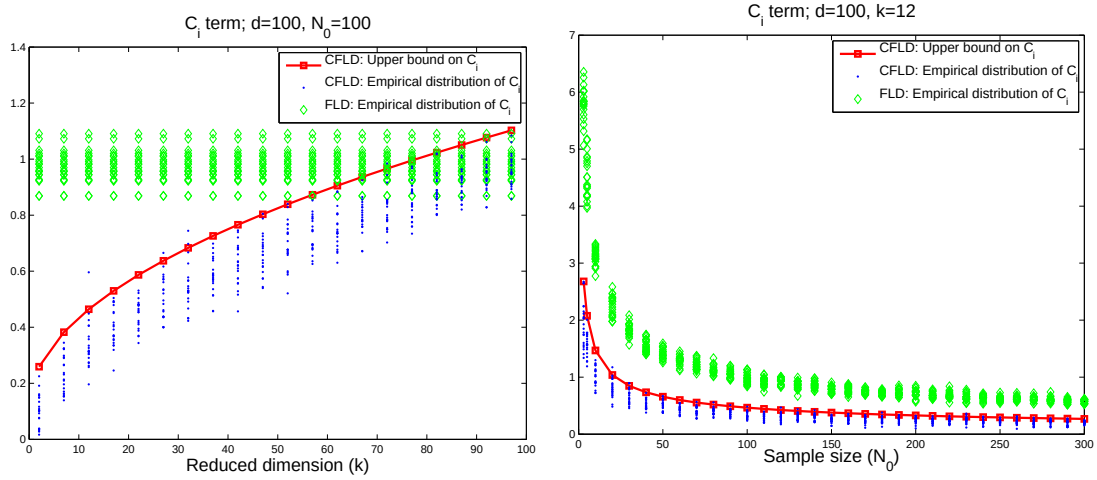
The terms C_i represent the mean estimation error for the i-th class.

For every instance of R, we have $(R\Sigma_i R^T)^{-\frac{1}{2}} R(\hat{\mu}_i - \mu_i) \sim \mathcal{N}(0, I_k/N_i)$. Hence, from Lemma 8 we have the following:

$$\|(R\Sigma_i R^T)^{-\frac{1}{2}} R(\mu_i - \hat{\mu}_i)\| \leqslant \sqrt{\frac{k}{N_i}} + \epsilon_{5,i}\sqrt{\frac{1}{N_i}} \tag{34}$$

w.p. $1 - \exp(-\epsilon_{5,i}^2/2)$.

Figure 5 shows numerical simulations illustrating our upper bound against the empirical behaviour of the distributions of a C_i term, and in comparison with the empirical behaviour of the corresponding term of dataspace FLD. The tightness of our bound is most apparent and we also see that random projection helps reduce the estimation errors. This is because estimation in lower dimensions is easier and needs less data points.



(a) (b)

Figure 5: Illustration of the upper bounds on a C_i term. Lower values are better. We see that random projection is beneficial. We also notice that our upper bound is tight and follows the empirical behaviour tightly.

3.5. Assembling the pieces

In the final step we simply put together the bounds from the previous subsections, and count up the number of distinct failure probabilities using the union bound, and putting $\epsilon_1 = \epsilon_{2i} = \epsilon_3 = \epsilon_4 = \epsilon_{5i} =: \epsilon$. The probability that the bound on A_i or $A_{\neg i}$ fails is bounded above by $4\exp(-\epsilon^2/2)$, the probability that the bound on B_i or $B_{\neg i}$ fails is bounded above by another $4\exp(-\epsilon^2/2)$ in the general case (or $2\exp(-\epsilon^2/2)$ in the special case when $\Sigma_0 = \Sigma_1$), and the probability that the bound on C_i or $C_{\neg i}$ fails is bounded above by a further $2\exp(-\epsilon^2/2)$. Thus, we get w.p. $1 - 10\exp(-\epsilon^2/2)$ the following:

$$\Pr_{x,y}[\hat{h}^R(Rx) \neq y | \mathcal{T}_N, N_0] \leqslant ...$$

$$\sum_{y=0}^{1} \pi_i \Phi \left(- \left[[\sqrt{k} - \epsilon]_+ \frac{\left[\sqrt{\|\mu_1 - \mu_0\|^2 + \frac{\mathrm{Tr}(N_1\Sigma_0 + N_0\Sigma_1)}{N_0 N_1}} - \epsilon \sqrt{\frac{\lambda_{\max}(\mathrm{Tr}(N_1\Sigma_0 + N_0\Sigma_1))}{N_0 N_1}} \right]_+}{\sqrt{\mathrm{Tr}(\Sigma_i)} + \sqrt{k\lambda_{\max}(\Sigma_i)} + \epsilon} g(\tilde{\kappa}_i) - \frac{\sqrt{k} + \epsilon}{\sqrt{N_i}} \right]_+ \right) \quad (35)$$

where $g(\tilde{\kappa}_i) = \frac{\sqrt{\tilde{\kappa}_i}}{1+\tilde{\kappa}_i}$, with $\tilde{\kappa}_i = \tilde{\kappa}_i(\epsilon)$ given by eq.(28) or by eq.(33).

This is the error for any random draw of the training set of size N where the class proportion N_0/N is fixed. To include the effects of a random proportion of class memberships according to $\Pr[y_i = 0] = \pi_0, i = 1, ...N$, we can use a Chernoff bound for Bernoulli

variables (easily derived from the Bernoulli moment generating function):

$$\forall \epsilon \in (0,1), \Pr[(1-\epsilon)\pi_0 N \leqslant N_0 \leqslant (1+\epsilon)\pi_0 N] > 1 - 2\exp(-\pi_0 N \epsilon^2/3) \tag{36}$$

Now, using the notations for $\alpha_0, \alpha_1, \beta_0$ and β_1 defined in the statement of the theorem, we get w.p. $1 - 2\exp(-\pi_0 N\epsilon^2/3)$ that,

$$N\beta_i \leq N_i \leq N\alpha_i. \tag{37}$$

Plugging back completes the proof. ■

4. Conclusions

We derived a non-asymptotic generalisation bound for the compressive Fisher discriminant classifier which is more complete than previous attempts, and is dimension free under mild assumptions. By decomposing the generalisation error we were able to disentangle the effects of random projection on various components of the error, pinpointing beneficial effects on misestimation and covariance misspecification, and a detrimental effect of reducing the class separation. We also gave an asymptotic bound as an immediate corollary of our result. A key technical ingredient in this analysis was to develop sharp dimension-free bounds on the largest and smallest eigenvalue of the compressive covariance by extending previous work using comparison inequalities for the suprema of Gaussian processes. In future work it will be of interest to investigate whether similarly sharp bounds could be derived for subGaussian random projection matrices.

References

W. Bian, D. Tao. Asymptotic Generalization Bound of Fisher's Linear Discriminant Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, 36(12), 2325-2337, 2014.

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.

S. Dasgupta. Learning mixtures of Gaussians. Proceedings of the 40-th Annual Symposium on Foundations of Computer Science (FOCS), volume 40, 1999, pp. 634-644.

K.R. Davidson, S.J. Szarek. Local operator theory, random matrices and Banach spaces. In Handbook of the geometry of Banach spaces, Vol. I, pp. 317-366, 2001.

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.

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), Journal of Machine Learning Research Workshop& Conference Proceedings 22: 337-345, 2012.

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.

R.A.Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188, 1936.

R.A. Horn, C.R. Johnson. Matrix analysis, CUP, 1985.

G.James and T. Hastie. Functional Linear Discriminant Analysis for Irregularly Sampled Curves. J. R. Stat. Soc. B 63 , 2001, pp. 533-550.

A. Kabán, R.J. Durrant. Dimension-Adaptive Bounds on Compressive FLD Classification. In Proc of the 24th International Conference on Algorithmic Learning Theory (ALT 2013), pp. 294-308, 2013.

M, Ledoux, M. Talagrand, M. Probability in Banach spaces. Berlin: Springer-Verlag, 1991.

G.J. McLachlan. Discriminant analysis and statistical pattern recognition. 1992. Wiley.

Iosif Pinelis. Expectation of Mahalanobis norm. mathoverflow.net/q/203309 (2015).

A Pourhabib, BK Mallick, Y Ding. Absent Data Generating Classifier for Imbalanced Class Sizes. J. of Machine Learning Research, 2015, to appear.

M. Rudelson, R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. arXiv:1306.2872 [math.PR] `http://arxiv.org/abs/1306.2872`

A.J. Scott, G.P.H. Styan. On a Separation Theorem for Generalized Eigenvalues and a Problem in the Analysis of Sample Surveys. Linear Algebra and its Applications, Vol. 70, October 1985, pp. 209-224.

H. Shin. An extension of Fisher's discriminant analysis for stochastic processes. Journal of Multivariate Analysis 99, 2008, pp. 1191-1216.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012. [arXiv:1011.3027, Aug 2010]

R. Vershynin. Spectral norm of products of random and deterministic matrices, Probability Theory and Related Fields 150 (2011), 471–509.