Random Projections as Regularizers: Learning a Linear Discriminant from Fewer Observations than Dimensions

Robert J. Durrant (bobd@waikato.ac.nz)
Department of Statistics, University of Waikato, Hamilton 3240, New Zealand.

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

May 28th 2014

Abstract. We prove theoretical guarantees for an averaging-ensemble of randomly projected Fisher Linear Discriminant classifiers, focusing on the case when there are fewer training observations than data dimensions.
The specific form and simplicity of this ensemble permits a direct and much more detailed analysis than existing generic tools in previous works. In particular, we are able to derive the exact form of the generalization error of our ensemble, conditional on the training set, and based on this we give theoretical guarantees which directly link the performance of the ensemble to that of the corresponding linear discriminant learned in the full data space. To the best of our knowledge these are the first theoretical results to prove such an explicit link for any classifier and classifier ensemble pair. Furthermore we show that the randomly projected ensemble is equivalent to implementing a sophisticated regularization scheme to the linear discriminant learned in the original data space and this prevents overfitting in conditions of small sample size where pseudo-inverse FLD learned in the data space is provably poor.
Our ensemble is learned from a set of randomly projected representations of the original high dimensional data and therefore for this approach data can be collected, stored and processed in such a compressed form.
We confirm our theoretical findings with experiments, and demonstrate the utility of our approach on several datasets from the bioinformatics domain and one very high dimensional dataset from the drug discovery domain, both settings in which fewer observations than dimensions are the norm.
A preliminary version of this work received the best paper award at the 5th Asian Conference on Machine Learning.

Keywords: Random Projections, Ensemble Learning, Linear Discriminant Analysis, Compressed Learning, Learning Theory

1. Introduction

Classification ensembles that use some form of randomization in the design of the base classifiers have a long and successful history in machine learning, especially in the case when there are fewer training observations than data dimensions. Common approaches include: Bagging [6]; random subspaces [30]; random forests [7].

Surprisingly, despite the well-known theoretical properties of random projections as dimension-reducing approximate isometries [11, 1] and empirical and theoretical studies demonstrating their usefulness when learning a *single* classifier (e.g. [20, 14]), results in the literature employing random projections to create weak learners for *ensemble* classification are sparse compared to results for other approaches such as bagging and random subspaces. On the other hand, given their appealing properties and tractability to analysis, random projections seem like a rather natural choice in this setting. Those empirical studies we could find on randomly-projected ensembles in the literature [22, 19, 44] all report good empirical performance from the ensemble, but none attempt a theoretical analysis. Indeed for all of the randomizing approaches mentioned above, despite a wealth of empirical evidence demonstrating the effectiveness of these ensembles, there are very few theoretical studies.

An important paper by [21] gives an approximate analytical model as a function of the ensemble size, applicable to linear combiners, which explains the variance reducing property of bagging. However, besides the inherent difficulties with the approach of bias-variance decomposition for classification problems [43], such analysis only serves to relate the performance of an ensemble to its members and [21] correctly point out that even for bagging, the simplest such approach and in use since at least 1996, there is 'no clear understanding yet of the conditions under which bagging outperforms an individual classifier [trained on the full original data set]'. They further state that, even with specific assumptions on the data distribution, such an analytical comparison would be a complex task. In other words, there is no clear understanding yet about when to use an ensemble vs. when to use one classifier.

Here we take a completely different approach to address this last open issue for a specific classifier ensemble: Focusing on an ensemble of randomly projected Fisher linear discriminant (RP-FLD) classifiers as our base learners, we leverage recent random matrix theoretic results to link the performance of the linearly combined ensemble to the corresponding classifier trained on the original data. In particular, we extend and simplify the work of [37] specifically for this classification setting, and one of our main contributions is to derive theoretical guarantees that directly link the performance of the randomly projected ensemble to the performance of Fisher linear discriminant (FLD) learned in the full data space. This theory is, however, not simply of abstract interest: We also show experimentally that the algorithm we analyze here is highly competitive with the state-of-the-art. Furthermore our algorithm has several practically desirable properties, amongst which are: Firstly, the individual ensemble members are learned in a very low-dimensional space from randomly-projected data, and so training data

can be collected, stored and processed entirely in this form. Secondly, our approach is fast – training on a single core typically has lower time complexity than learning a regularized FLD in the data space, while for classification the time complexity is the same as the data space FLD. Thirdly, parallel implementation of our approach is straightforward since, both for training and classification, the individual ensemble members can be run on separate cores. Finally, our approach returns an inverse covariance matrix estimate for the full d-dimensional data space, the entries of which are interpretable as conditional correlations; this may be useful in a wide range of settings.

Our randomly projected ensemble approach can be viewed as a generalization of bagged ensembles, in the sense that here we generate multiple instances of training data by projecting a training set of size N onto a subspace drawn uniformly at random with replacement from the data space, whereas in bagging one generates instances of training data by drawing N' training examples uniformly with replacement from a training set of size $N \geqslant N'$. However, in this setting, an obvious advantage of our approach over bagging is that it is able to repair the rank deficiency of the sample covariance matrix we need to invert in order to build the classifier. In particular, we show that when there are fewer observations than dimensions our ensemble implements a data space FLD with a sophisticated covariance regularization scheme (parametrized by an integer parameter) that subsumes a combination of several previous regularization schemes. In order to see the clear structural links between our ensemble and its data space counterpart we develop our theory in a random matrix theoretic setting. We avoid a bias-variance decomposition approach since, in common with the analysis of [43], a key property of our ensemble is that its effect is not simply to reduce the variance of a biased classifier.

The structure of the remainder of the paper is as follows: We give some brief background and describe the randomly projected FLD classifier ensemble. Next, we present theoretical findings that give insight into how this ensemble behaves. We continue by presenting extensive experiments on real datasets from the bioinformatic domain where FLD (and variants) are a popular classifier choice even though often restricted to a diagonal covariance choice because of high dimensionality and data scarcity [25, 13]. We further present experimental results on a 100,000-dimensional drug discovery dataset, that is from another problem domain where the small sample size problem typically arises. Our experiments suggest that in practice, when the number of training examples is less than the number of data dimensions, our ensemble approach outperforms the traditional FLD in the data space both in terms of prediction performance and computation time. Finally, we

summarize and discuss possible future directions for this and similar approaches.

2. Preliminaries

We consider a binary classification problem in which we observe N i.i.d examples of labelled training data $\mathcal{T}_N = \{(x_i, y_i) : x_i \in \mathbb{R}^d, y_i \in \{0,1\}\}_{i=1}^N$ where $(x_i, y_i) \overset{i.i.d}{\sim} \mathcal{D}_{x,y}$. We are interested in comparing the performance of a randomly-projected ensemble classifier working in the projected space $\mathbb{R}^k$, $k \ll d$, to the performance achieved by the corresponding classifier working in the data space $\mathbb{R}^d$. We will consider Fisher's linear discriminant classifier working in both of these settings since FLD is a popular and widely used linear classifier (in the data space setting) and yet it is simple enough to analyse in detail.
The decision rule for FLD learned from training data is given by:

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

where $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\Sigma}$ are maximum likelihood (ML) estimates of the class-conditional means and (shared) covariance matrix respectively, and $\mathbf{1}(\cdot)$ is the indicator function which returns 1 if its argument is true and 0 otherwise. In the setting considered here we assume that $N \ll d$. Hence, $\hat{\Sigma}$ will be singular and so one can either pseudo-invert or regularize $\hat{\Sigma}$ to obtain a working decision rule; both approaches are used in practice [42].
To construct the randomly projected ensemble, we choose the number of ensemble members M and the projection dimensionality k, and generate M random matrices $R \in \mathcal{M}_{k\times d}$ with i.i.d entries $r_{ij} \sim \mathcal{N}(0, \sigma^2)$ each. We can take $\sigma^2 = 1$ without loss of generality. Such matrices are called random projection matrices in the literature [4, 1]. Pre-multiplying the data with one of the matrices R maps the training examples to a random k-dimensional subspace of the data space $\mathbb{R}^d$ and for each instance of R we learn a single FLD classifier in this subspace. By linearity of expectation and of the projection operator, the decision rule for a single randomly projected classifier is then given by:

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

For an ensemble, various different combination rules can be applied. The most common choices include majority voting (when there is an

odd number of classifiers in the ensemble) and linear combination [8]. We want to make the most of the weak learners' confidence estimates so we choose to employ the averaged linear decisions of M base learners as our combination rule which gives the following ensemble decision:

$$\hat{h}_{ens}(x_q) := \mathbf{1}\left\{\frac{1}{M}\sum_{i=1}^{M}(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i\hat{\Sigma}R_i^T\right)^{-1} R_i\left(x_q - \frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right) > 0\right\}$$

Our algorithm is therefore very simple: We learn M FLD classifiers from M different instances of randomly projected data, average their outputs and take the sign of this average as the ensemble decision. This combination rule is called 'voting' in the ensemble literature but, to avoid any possible confusion with majority voting, we shall refer to it as 'RP averaging'; it does not require the number of classifiers in the ensemble to be odd for good generalization and, as we shall see, it also has the advantage of analytical tractability.

We commence our theoretical analysis of this algorithm by examining the expected performance of the RP-FLD ensemble when the training set is fixed, which is central to linking the ensemble and data space classifiers, and then later in Theorem 3.2 we will consider random instantiations of the training set.

To begin, observe that by the law of large numbers the LHS of the argument of the decision rule of the ensemble converges to the following:

$$\begin{aligned}\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}&(\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i\left(x_q - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)\\ &= (\hat{\mu}_1 - \hat{\mu}_0)^T\mathrm{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]\left(x_q - \frac{\hat{\mu}_1+\hat{\mu}_0}{2}\right)\end{aligned} \tag{2.1}$$

provided that this limit exists. If ρ is the rank of $\hat{\Sigma}$, then it will turn out that for $R \in \mathcal{M}_{k\times d}$ having i.i.d zero-mean Gaussian entries $r_{ij} \sim \mathcal{N}(0,1)$, if $k \in \{1,...,\rho-2\}$ then this expectation is indeed defined for each entry. From equation (2.1) we see that, for a fixed training set, in order to quantify the error of the ensemble it is enough to consider the expectation (w.r.t random matrices R):

$$\mathrm{E}\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right] \tag{2.2}$$

Before continuing, we should note that for the case $k \in \{1,...,\rho-2\}$ [37] provide a procedure to compute this expectation exactly. However we are more interested in how this expectation relates to characteristics of

the maximum likelihood estimate of the sample covariance $\hat{\Sigma}$, since we shall see in theorem 3.2 that improving the conditioning of this matrix has a direct impact on the generalization error of the FLD classifier. Our approach and proof techniques are therefore very different to those followed by [37], specifically we bound this expectation from both sides in the positive semi-definite ordering in order to provide an estimate of the extreme eigenvalues of the inverse covariance matrix implemented by our ensemble.

3. Theory

Our main theoretical results are the following three theorems: The first characterizes the regularization effect of our ensemble, while the second bounds the generalization error of the ensemble for an arbitrary training set of size N in the case of multivariate Gaussian class-conditional distributions with shared covariance. The third is a finite sample generalization of the negative result of [5] showing that when the data dimension d is large compared to the rank of $\hat{\Sigma}$ (which is a function of the sample size) then, with high probability, pseudoinverted FLD performs poorly.

Theorem 3.1 (Regularization)**.** *Let $\hat{\Sigma} \in \mathcal{M}_{d\times d}$ be a symmetric positive semi-definite matrix with rank $\rho \in \{3, ..., d-1\}$, and denote by $\lambda_{\max}(\hat{\Sigma}), \lambda_{\min\neq 0}(\hat{\Sigma}) > 0$ its greatest and least non-zero eigenvalues. Let $k < \rho - 1$ be a positive integer, and let $R \in \mathcal{M}_{k\times d}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. Let $S^{-1} := E\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$, and denote by $\kappa(S^{-1})$ its condition number, $\kappa(S^{-1}) = \lambda_{\max}(S^{-1})/\lambda_{\min}(S^{-1})$. Then:*

$$\kappa(S^{-1}) \leqslant \frac{\rho}{\rho - k - 1} \cdot \frac{\lambda_{\max}(\hat{\Sigma})}{\lambda_{\min\neq 0}(\hat{\Sigma})}$$

This theorem implies that for a large enough ensemble the condition number of the sum of random matrices $\frac{1}{M}\sum_{i=1}^{M} R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i$ is bounded. Of course, any one of these summands $R_i^T\left(R_i\hat{\Sigma}R_i^T\right)^{-1}R_i$ is singular by construction. On the other hand if we look at the decision rule of a single randomly projected classifier in the k-dimensional space,

$$\hat{h}_R(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)R^T(R\hat{\Sigma}R^T)^{-1}R\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\} \tag{3.1}$$

we have for all $z \neq 0$, $Rz \neq 0$ almost surely, and $R\hat{\Sigma}R^T$ is full rank almost surely – therefore with probability 1 the k-dimensional system in (3.1) is well-posed.

The significance of this theorem from a generalization error analysis point of view stems from the fact that the rank deficient maximum-likelihood covariance estimate has unbounded condition number and, as we see below in theorem 3.2, (an upper bound on) the generalization error of FLD increases as a function of the condition number of the covariance estimate employed. In turn, the bound given in our theorem 3.1 depends on the extreme *non-zero* eigenvalues of $\hat{\Sigma}$, its rank[1] ρ, and the subspace dimension k, which are all finite for any particular training set instance. We should also note that the subspace dimension k is a parameter that we can choose, and in what follows k therefore acts as the integer regularization parameter in our setting.

Theorem 3.2 (Tail bound on generalization error of the converged ensemble)**.** *Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ be a set of training data of size $N = N_0 + N_1$, subject to $N < d$ and $N_y > 1\ \forall y$. Let x_q be a query point with Gaussian class-conditionals $x_q|(y_q = y) \sim \mathcal{N}(\mu_y, \Sigma)$, and let $\Pr\{y_q = y\} = \pi_y$. Let ρ be the rank of the maximum likelihood estimate of the covariance matrix and let $k < \rho - 1$ be a positive integer. Then for any $\delta \in (0, 1)$ and any training set of size N, the generalization error of the converged ensemble of randomly projected FLD classifiers is upper-bounded with probability at least $1 - \delta$ by the following:*

$$\Pr_{x_q, y_q}(\hat{h}_{ens}(x_q) \neq y_q) \leqslant \sum_{y=0}^{1} \pi_y \Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right)\right.\right.$$
$$\ldots \times \left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 + \frac{dN}{N_0 N_1}} - \sqrt{\frac{2N}{N_0 N_1}\log\frac{5}{\delta}}\right]_+$$
$$\left.\left. \ldots - \sqrt{\frac{d}{N_y}}\left(1 + \sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right)$$

Where Φ is the c.d.f of the standard Gaussian, $\bar{\kappa}(\epsilon)$ is a high probability (w.r.t. the random draw of training set) upper bound on the condition number of $\Sigma\hat{S}^{-1}$ given by eq. (4.17) and $g(\cdot)$ is the function $g(a) := \frac{\sqrt{a}}{1+a}$.

The principal terms in this bound are: (i) The function $g : [1, \infty) \to (0, \frac{1}{2}]$ which is a decreasing function of its argument and here captures the effect of the mismatch between the estimated model covariance matrix $\hat{S}^{-1}$ and the true class-conditional covariance Σ, via a high-probability upper bound on the condition number of $\hat{S}^{-1}\Sigma$; (ii) The

[1] In the setting considered here we typically have $\rho = N - 2$

Mahalanobis distance between the two class centres which captures the fact that the better separated the classes are the smaller the generalization error should be; and (iii) antagonistic terms involving the sample size (N) and the number of training examples in each class (N_0, N_1), which capture the effect of class (im)balance – the more evenly the training data are split, the tighter the bound.
We note that a bound on generalization error with similar behaviour can be obtained for the much larger family of sub-Gaussian distributions, or when the true class-conditional covariance matrices are taken to be different (see e.g. [14, 16]). Therefore the distributional assumptions on Theorem 3.2 are not crucial.

Theorem 3.3 (High probability lower bound on generalization error of pseudoinverted FLD)**.** *For any $\delta \in (0,1)$, and any data set of size $N_0 + N_1 = N$, assuming Gaussian classes with shared covariance and $\kappa(\Sigma) < \infty$, the generalization error of pseudo-inverted FLD is lower-bounded with probability at least $1-\delta$ over the random draws of training set by:*

$$\Pr(\hat{h}_+(x_q) \neq y_q) \geqslant \Phi \left(-\frac{1}{2}\sqrt{1+\sqrt{\frac{8}{N}\log\frac{2}{\delta}}} \right.$$
$$\ldots \times \left(1+\sqrt{\frac{2\lambda_{\max}(\Sigma)\log(2/\delta)}{Tr(\Sigma)+\|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}} \right)$$
$$\left. \ldots \times \sqrt{\frac{\rho}{d}\frac{\|\mu_1-\mu_0\|^2+Tr(\Sigma)\frac{N}{N_0N_1}}{\lambda_{\min}(\Sigma)}} \right)$$

Where Φ is the c.d.f of the standard Gaussian.

It is interesting to notice that this lower bound depends on the rank of the covariance estimate, not on its fit to the true covariance Σ. Note in particular that when $N \ll d$ our lower bound explains the bad performance of pseudo-inverted FLD since ρ, the rank of $\hat{\Sigma}$, is at most $\min\{N-2, d\}$ and the lower bound of theorem 3.3 becomes tighter as ρ/d decreases. Allowing the dimensionality d to be large, as in [5], so that $\rho/d \to 0$, this fraction goes to 0 which means the lower bound of Theorem 3.3 converges to $\Phi(0) = 1/2$ – in other words random guessing.

4. Proofs

4.1. Proof of Theorem 3.1

Estimating the condition number of $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ is the key result underpinning our generalization error results. We will make use of the following two easy, but useful, lemmas:

Lemma 4.1 (Unitary invariance). *Let $R \in \mathcal{M}_{k\times d}$ with $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$. Let $\hat{\Sigma}$ be any symmetric positive semi-definite matrix, and let $\hat{U}$ be a unitary matrix such that $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$, where $\hat{\Lambda}$ is a diagonal matrix with the eigenvalues of $\hat{\Sigma}$ in descending order along the diagonal. Then:*

$$E\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right] = \hat{U}E\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]\hat{U}^T$$

Lemma 4.2 (Expected preservation of eigenvectors). *Let $\hat{\Lambda}$ be a diagonal matrix, then $E\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ is a diagonal matrix.*
Furthermore, if $\hat{U}$ diagonalizes $\hat{\Sigma}$ as $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$, then $\hat{U}$ also diagonalizes $E\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$.

We omit the proofs which are straightforward and can be found in [37].

Now, it follows from lemmas 4.1 and 4.2 that at convergence our ensemble preserves the eigenvectors of $\hat{\Sigma}$, and so we only need to consider the diagonal entries (i.e. the eigenvalues) of $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$, which we now do. To fix ideas we will look first at the case $k = 1$, when we are projecting the high dimensional data on to a single line for each classifier in the ensemble. In this case the i-th diagonal element of $\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]$ is $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho}\lambda_j r_j^2}\right]$, where r_i is the i-th entry of the single row matrix R. This can be upper and lower bounded as:

$$\frac{1}{\lambda_{\max}}\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] \leqslant \mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho}\lambda_j r_j^2}\right] \leqslant \frac{1}{\lambda_{\min\neq 0}}\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right]$$

where $\lambda_{\min\neq 0}$ denotes the smallest nonzero eigenvalue of $\hat{\Lambda}$ (and of $\hat{\Sigma}$), and $\lambda_{\max}$ its largest eigenvalue.
Recall that as a result of lemmas 4.1 and 4.2 we only need consider the diagonal entries of this expectation as the off-diagonal terms are known

to be zero.
Now, we evaluate the remaining expectation. There are two cases: If $i > \rho$ then r_i is independent from the denominator and we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[r_i^2\right] \mathrm{E}\left[1/\sum_{j=1}^{\rho} r_j^2\right] = \frac{1}{\rho-2}$, where we used the expectation of the inverse-χ^2 with ρ degrees of freedom, and the fact that $\mathrm{E}\left[r_i^2\right] = 1$. When $i \leqslant \rho$, then in turn we have $\mathrm{E}\left[\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right] = \mathrm{E}\left[\frac{r_i^2}{\|r\|^2}\right] = \frac{1}{\rho}$. That is,

$$\mathrm{E}\left[\mathrm{diag}\left(\frac{r_i^2}{\sum_{j=1}^{\rho} r_j^2}\right)\right] = \left[\begin{array}{c|c} \frac{1}{\rho} I_\rho & 0 \\ \hline 0 & \frac{1}{\rho-2} I_{d-\rho} \end{array}\right]$$

and so $\mathrm{E}\left[R^T \left(R\hat{\Lambda}R^T\right)^{-1} R\right]$ is full rank, hence invertible. Its inverse may be seen as a regularized covariance estimate in the data space, and its condition number, κ, is upper bounded by:

$$\kappa \leqslant \frac{\rho}{\rho - 2} \cdot \frac{\lambda_{\max}}{\lambda_{\min \neq 0}}$$

whereas in the setting $N < d$ the ML covariance estimate has unbounded condition number.
For the general $k < \rho - 1$ case we write R as a concatenation of two matrices $R = [P, S]$ where P is $k \times \rho$ and S is $k \times (d - \rho)$, so that $\mathrm{E}\left[R^T \left(R\hat{\Lambda}R^T\right)^{-1} R\right]$ can be decomposed as two diagonal blocks:

$$\left[\begin{array}{c|c} \mathrm{E}[P^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} P] & 0 \\ \hline 0 & \mathrm{E}[S^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} S] \end{array}\right]$$

Where here in $P\underline{\hat{\Lambda}}P^T$ we use $\underline{\hat{\Lambda}}$ to denote the $\rho \times \rho$ positive definite upper block of the positive semi-definite matrix $\hat{\Lambda}$. Now, rewrite the upper block to orthonormalize P as the following: $\mathrm{E}[P^T \left(P\underline{\hat{\Lambda}}P^T\right)^{-1} P] =$

$$\mathrm{E}[P^T (PP^T)^{-\frac{1}{2}} \left((PP^T)^{-\frac{1}{2}} P\underline{\hat{\Lambda}}P^T (PP^T)^{-\frac{1}{2}}\right)^{-1} (PP^T)^{-\frac{1}{2}} P]$$

Denoting by P_i the i-th column of P, we can write and bound the i-th diagonal element as:

$$\mathrm{E}[P_i^T (PP^T)^{-\frac{1}{2}} \left((PP^T)^{-\frac{1}{2}} P\underline{\hat{\Lambda}}P^T (PP^T)^{-\frac{1}{2}}\right)^{-1} (PP^T)^{-\frac{1}{2}} P_i]$$

$$\leqslant \mathrm{E}\left[\frac{P_i^T(PP^T)^{-1}P_i}{\lambda_{\min}((PP^T)^{-\frac{1}{2}}P\hat{\underline{\Lambda}}P^T(PP^T)^{-\frac{1}{2}})}\right]$$
$$\leqslant \mathrm{E}\left[\frac{P_i^T(PP^T)^{-1}P_i}{\lambda_{\min\neq 0}}\right]$$

with $\lambda_{\min\neq 0}$ the smallest non-zero eigenvalue of $\hat{\Lambda}$ as before, and where we used the Rayleigh quotient and the Poincaré separation theorem respectively (e.g. [31] Thm 4.2.2, Corr 4.3.16). This holds for all i, so then replacing we have:

$$\mathrm{E}[P^T(PP^T)^{-1}P]/\lambda_{\min\neq 0} \succcurlyeq \mathrm{E}\left[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P\right] \tag{4.1}$$

where $A \succcurlyeq B$ denotes $A-B$ is positive semi-definite. Now the remaining expectation can be evaluated using the expectation of the ρ-dimensional Wishart matrix P^TP with k degrees of freedom:

$$\mathrm{E}[P^T(PP^T)^{-1}P] = \mathrm{E}[P^TP]/\rho = \frac{k}{\rho}\cdot I_\rho$$

Similarly to equation (4.1) we can also show that:

$$\mathrm{E}\left[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P\right] \succcurlyeq \mathrm{E}[P^T\left(PP^T\right)^{-1}P]/\lambda_{\max} \tag{4.2}$$

in much the same way. Put together, the diagonal elements in the upper block are all in the interval:

$$\left[\frac{1}{\lambda_{\max}}\frac{k}{\rho}, \frac{1}{\lambda_{\min\neq 0}}\frac{k}{\rho}\right]$$

Hence, we see that in this upper block the condition number is reduced in comparison to that of $\hat{\Lambda}$ in its column space.

$$\frac{\lambda_{\max}(\mathrm{E}[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P])}{\lambda_{\min}(\mathrm{E}[P^T(P\hat{\underline{\Lambda}}P^T)^{-1}P])} \leqslant \frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min\neq 0}(\hat{\Lambda})}$$

That is, in the range of $\hat{\Sigma}$, the ensemble has the effect of a shrinkage regularizer [34]. Next, we consider its effect in the null space of $\hat{\Sigma}$. The lower block is $\mathrm{E}\left[S^T(P\hat{\underline{\Lambda}}P^T)^{-1}S\right] = \mathrm{Tr}\left(\mathrm{E}\left[(P\hat{\underline{\Lambda}}P^T)^{-1}\right]\right)\cdot I_{d-\rho}$ since S is independent of P. We again rewrite this to orthonormalize P. Going through similar steps, we obtain: $\mathrm{Tr}\left(\mathrm{E}\left[(P\hat{\underline{\Lambda}}P^T)^{-1}\right]\right) =$

$$\mathrm{Tr}\left(\mathrm{E}\left[\left(PP^T\right)^{-\frac{1}{2}}\left(\left(PP^T\right)^{-\frac{1}{2}}P\hat{\underline{\Lambda}}P^T\left(PP^T\right)^{-\frac{1}{2}}\right)^{-1}\left(PP^T\right)^{-\frac{1}{2}}\right]\right)$$

$$\leqslant \frac{\mathrm{Tr}\left(\mathrm{E}\left[\left(PP^T\right)^{-1}\right]\right)}{\lambda_{\min\neq 0}} = \frac{k}{\rho-k-1}\cdot\frac{1}{\lambda_{\min\neq 0}}$$

where we used the expectation of the inverse Wishart. Likewise,

$$\mathrm{Tr}\left(\mathrm{E}\left[\left(P\hat{\underline{\Lambda}}P^T\right)^{-1}\right]\right) \geqslant \frac{k}{\rho-k-1}\cdot\frac{1}{\lambda_{\max}} \tag{4.3}$$

Hence, the lower block is a multiple of $I_{d-\rho}$ with the coefficient in the interval:

$$\left[\frac{k}{\rho-k-1}\frac{1}{\lambda_{\max}}, \frac{k}{\rho-k-1}\frac{1}{\lambda_{\min\neq 0}}\right]$$

That is, in the null space of $\hat{\Sigma}$ the ensemble acts as a ridge regularizer [29].

Putting everything together, the condition number of the covariance (or inverse covariance) estimate is upper bounded by:

$$\kappa \leqslant \frac{\rho}{\rho-k-1}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{4.4}$$

which we see reduces to equation (4.1) when $k=1$. □

4.2. Proof of Theorem 3.2

Traditionally ensemble methods are regarded as 'meta-learning' approaches and although bounds exist (e.g. [33]) there are, to the best of our knowledge, no results giving the exact analytical form of the generalization error of any particular ensemble. Indeed, in general it is not analytically tractable to evaluate the generalization error exactly, so one can only derive bounds. Because we deal with a particular ensemble of FLD classifiers we are able to derive the exact generalization error of the ensemble in the case of Gaussian classes with shared covariance Σ, the setting in which FLD is Bayes' optimal. This allows us to explicitly connect the performance of the ensemble to its data space analogue. As noted earlier, an upper bound on generalization error with similar behaviour can be derived for the much larger class of sub-Gaussian distributions (see e.g. [14, 16]), therefore this Gaussianity assumption is not crucial.

We proceed in two steps: (1) Obtain the generalization error of the ensemble conditional on a fixed training set; (2) Bound the deviation of this error caused by a random draw of a training set.

4.2.1. *Generalization error of the ensemble for a fixed training set*

For a fixed training set, the generalization error is given by the following lemma:

Lemma 4.3 (Exact generalization error with Gaussian classes)**.** *Let* $x_q|(y_q = y) \sim \mathcal{N}(\mu_y, \Sigma)$*, where* $\Sigma \in \mathcal{M}_{d\times d}$ *is a full rank covariance matrix, and let* $\pi_y := Pr\{y_q = y\}$*. Let* $R \in \mathcal{M}_{k\times d}$ *be a random projection matrix with i.i.d. zero-mean Gaussian entries and denote* $\hat{S}^{-1} := E_R\left[R^T\left(R\hat{\Sigma}R^T\right)^{-1}R\right]$*. Then the exact generalization error of the converged randomly projected ensemble classifier* (2.1) *is given by* $\Pr_{(x_q,y_q)}\{\hat{h}_{ens}(x_q) \neq y_q\} =$

$$\sum_{y=0}^{1}\pi_y\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T\hat{S}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_y)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right) \tag{4.5}$$

Where Φ *is the c.d.f of the standard Gaussian.*

The proof of this lemma is similar in spirit to the one for a single FLD in [40]. For completeness we give it below.

Proof of Lemma 4.3

Without loss of generality let x_q have label 0. By assumption the classes have Gaussian distribution $\mathcal{N}(\mu_y, \Sigma)$ so then the probability that x_q is misclassified by the converged ensemble is given by:

$$\Pr_{x_q|y_q=0}\left\{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)>0\right\} \tag{4.6}$$

Define $a^T := (\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}$ and observe that if $x_q \sim \mathcal{N}(\mu_0, \Sigma)$ then:

$$\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right), \Sigma\right)$$

and so:

$$a^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right) \sim \mathcal{N}\left(a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right), a^T\Sigma a\right)$$

which is a univariate Gaussian. Therefore:

$$\frac{a^T\left(x_q-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)-a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}} \sim \mathcal{N}(0,1)$$

Hence, for the query point x_q we have the probability (4.6) is given by:

$$\begin{aligned}&\Phi\left(\frac{a^T\left(\mu_0-\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)}{\sqrt{a^T\Sigma a}}\right)\\ &=\Phi\left(-\tfrac{1}{2}\frac{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right)\end{aligned} \tag{4.7}$$

where Φ is the c.d.f of the standard Gaussian.
A similar argument deals with the case when x_q belongs to class 1, and applying the law of total probability completes the proof. □
Indeed equation (4.5) has the same form as the error of the data space FLD (See [5, 40] for example.) and the converged ensemble, inspected in the original data space, produces exactly the same mean estimates and covariance matrix eigenvector estimates as FLD working on the original data set. However it has different eigenvalue estimates that result from the sophisticated regularization scheme that we analyzed in section 4.1.

4.2.2. *Tail bound on the generalization error of the ensemble.*
The previous section gave the exact generalization error of our ensemble conditional on a given training set. In this section our goal is to derive an upper bound with high probability on the ensemble generalization error w.r.t. random draws of the training set.

We will use the following concentration lemma:

Lemma 4.4 (Concentration bound on exponential random variables)**.** *Let X be a Gaussian random vector in $\mathbb{R}^d$ with mean vector $E[X] = \mu$ and covariance matrix Σ. Let $\epsilon > 0$. Then:*

$$\begin{aligned} &Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(Tr(\Sigma) + \|\mu\|^2\right)\right\} \\ &\leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon} - 1\right)^2\right) \end{aligned} \tag{4.8}$$

Furthermore, if $\epsilon \in (0,1)$:

$$\begin{aligned} &Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(Tr(\Sigma) + \|\mu\|^2\right)\right\} \\ &\leqslant \exp\left(-\frac{Tr(\Sigma) + \|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \end{aligned} \tag{4.9}$$

The proof, which follows immediately from the more general result we give in [15], is given in appendix A for completeness. Now we can bound the generalization error of the RP-FLD ensemble. We begin by decomposing the numerator of the generalization error term (for a single class) obtained in lemma 4.3 as follows:

$$\begin{aligned} &\left(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0\right)^T \hat{S}^{-1}\left(\hat{\mu}_1 - \hat{\mu}_0\right) \\ =\ &\left(\hat{\mu}_1 - \hat{\mu}_0\right)^T \hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) + 2\left(\hat{\mu}_0 - \mu_0\right)^T \hat{S}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) \end{aligned}$$

Using this decomposition we can rewrite the argument of the first term in lemma 4.3 in the following form:

$$\Phi\left(-\frac{1}{2}[A-B]\right)$$

Where:

$$A = \frac{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} \Sigma \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}}$$

and:

$$B = \frac{2\,(\mu_0 - \hat{\mu}_0)^T \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{S}^{-1} \Sigma \hat{S}^{-1} (\hat{\mu}_1 - \hat{\mu}_0)}}$$

We will lower bound A and upper bound B with high probability over the random draw of training set in order to bound the whole term from above with high probability and, since Φ is monotonic increasing in its argument, this will give the upper bound on generalization error.

Lower-bounding the term A
Applying the Kantorovich inequality (e.g. [31] Thm 7.4.41), A is lower bounded by:

$$\|\Sigma^{-\frac{1}{2}} (\hat{\mu}_1 - \hat{\mu}_0)\| \cdot \frac{2\sqrt{\kappa(\hat{S}^{-\frac{1}{2}} \Sigma \hat{S}^{-\frac{1}{2}})}}{1 + \kappa(\hat{S}^{-\frac{1}{2}} \Sigma \hat{S}^{-\frac{1}{2}})} \tag{4.10}$$

where $\kappa(H) := \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}$ denotes the condition number of the matrix H.

Next, since $\Sigma^{-\frac{1}{2}}\hat{\mu}_1$ and $\Sigma^{-\frac{1}{2}}\hat{\mu}_0$ are independent with $\Sigma^{-\frac{1}{2}}\hat{\mu}_y \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}\mu_y, I_d/N_y)$, we have $\Sigma^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0) \sim \mathcal{N}(\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0), N/(N_0 N_1) \cdot I_d)$.

Applying the second concentration bound of lemma 4.4, (4.9), we have:

$$\|\Sigma^{-\frac{1}{2}}(\hat{\mu}_1 - \hat{\mu}_0)\| \geqslant \sqrt{(1-\epsilon)\left(\frac{d \cdot N}{N_0 N_1} + \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2\right)} \tag{4.11}$$

with probability at least:

$$1 - \exp\left(-\frac{d + \|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|^2 N_0 N_1 / N}{2}\left(\sqrt{1-\epsilon} - 1\right)^2\right) \tag{4.12}$$

To complete the bounding of the term A, we denote $g(a) := \frac{\sqrt{a}}{1+a}$, and observe that this is a monotonic decreasing function on $[1, \infty)$. So, replacing a with the condition number $\kappa(\hat{S}^{-\frac{1}{2}} \Sigma \hat{S}^{-\frac{1}{2}}) \in [1, \infty)$ we

need to upper bound this condition number in order to lower bound g. Denoting this upper bound by $\bar{\kappa}$, which will be quantified in lemma 4.5, then the term A is lower bounded with high probability by:

$$A \geqslant 2g(\bar{\kappa})\sqrt{(1-\epsilon)\left(\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2 + \frac{d\cdot N}{N_0 N_1}\right)} \tag{4.13}$$

Upper-bounding the term B
We can rewrite B by inserting $\Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}} = I_d$, and using Cauchy-Schwarz in the numerator to give:

$$B \leqslant \frac{2\|\Sigma^{-\frac{1}{2}}(\mu_0-\hat{\mu}_0)\|\cdot\|\Sigma^{\frac{1}{2}}\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\|}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}} \tag{4.14}$$

After cancellation, this simplifies to:

$$= 2\|\Sigma^{-\frac{1}{2}}(\mu_0-\hat{\mu}_0)\| \tag{4.15}$$

and so by Lemma 4.4, (4.8), we have:

$$B \;\leqslant\; 2\sqrt{(1+\epsilon)d/N_0} \tag{4.16}$$

with probability at least $1-\exp(-\frac{d}{2}(\sqrt{1+\epsilon}-1)^2)$.

To bound the condition number $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})$ with high probability we need the following additional lemma:

Lemma 4.5 (Upper bound on $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})$)**.** *Under the conditions of theorem 3.2 we have,* $\forall\epsilon > 0$*:*

$$\begin{aligned}\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}}) &= \frac{\lambda_{\max}(\Sigma^{\frac{1}{2}}\cdot E_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\cdot\Sigma^{\frac{1}{2}})}{\lambda_{\min}(\Sigma^{\frac{1}{2}}\cdot E_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\cdot\Sigma^{\frac{1}{2}})} \\ &\leqslant \frac{(\sqrt{N-2}+\sqrt{d}+\epsilon)^2(1+\rho/k\cdot\kappa(\Sigma))}{(\sqrt{N-2}-\sqrt{k}-\epsilon)^2} \\ &=: \bar{\kappa}(\epsilon)\end{aligned} \tag{4.17}$$

with probability at least $1-2\exp(-\epsilon^2/2)$.

4.2.3. *Proof of lemma 4.5*
To upper bound the condition number $\kappa(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}})$ with high probability, we derive high-probability upper and lower bounds on (respectively) the greatest and least eigenvalues of its argument. We will make use of the following result, Eq. (2.3) from [47]:

Lemma 4.6 (Singular Values of Wishart Matrices [47]). *Let R be a $k \times d$ matrix with i.i.d $\mathcal{N}(0,1)$ entries. Then for all $\epsilon > 0$ with probability at least $1 - 2\exp(-\epsilon^2/2)$ we have:*

$$\sqrt{d} - \sqrt{k} - \epsilon \leqslant s_{\min}(R) \leqslant s_{\max}(R) \leqslant \sqrt{d} + \sqrt{k} + \epsilon \tag{4.18}$$

Upper-bound on largest eigenvalue
By Jensen's inequality, and noting that $\lambda_{\max}(\cdot)$ is a convex function, we have:

$$\begin{aligned}
& \lambda_{\max}(\hat{S}^{-\frac{1}{2}}\Sigma\hat{S}^{-\frac{1}{2}}) \\
= \; & \lambda_{\max}(\Sigma^{\frac{1}{2}}\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R]\Sigma^{\frac{1}{2}}) \\
\leqslant \; & \mathrm{E}_R[\lambda_{\max}(\Sigma^{\frac{1}{2}}R^T(R\hat{\Sigma}R^T)^{-1}R\Sigma^{\frac{1}{2}})] \\
= \; & \mathrm{E}_R[\lambda_{\max}((R\hat{\Sigma}R^T)^{-1}R\Sigma R^T] \\
= \; & \mathrm{E}_R[\lambda_{\max}((R\Sigma R^T)^{\frac{1}{2}}(R\hat{\Sigma}R^T)^{-1}(R\Sigma R^T)^{\frac{1}{2}})] \\
= \; & \mathrm{E}_R\left[\frac{1}{\lambda_{\min}((R\Sigma R^T)^{-\frac{1}{2}}R\hat{\Sigma}R^T(R\Sigma R^T)^{-\frac{1}{2}})}\right] \\
\leqslant \; & \frac{N}{(\sqrt{N-2} - \sqrt{k} - \epsilon)^2}
\end{aligned}$$

with probability at least $1 - \exp(-\epsilon^2/2), \forall \epsilon > 0$, where throughout we use the fact that the non-zero eigenvalues of AB are the same as non-zero eigenvalues of BA, in the second to last step we used the fact that for invertible matrices A we have $\lambda_{\max}(A) = 1/\lambda_{\min}(A^{-1})$, and in the last step we used that for any particular full row-rank matrix R, $(R\Sigma R^T)^{-\frac{1}{2}}R\hat{\Sigma}R^T(R\Sigma R^T)^{-\frac{1}{2}}$ (regarded as a function of the training set and therefore $\hat{\Sigma}$ is the random variable) is distributed as a k-dimensional Wishart with $N-2$ degrees of freedom and scale matrix I_k (e.g. [36] Corr. 3.4.1.2), and we then used the high probability lower-bound for the smallest eigenvalue of such a matrix, lemma 4.6.

Lower-bound on smallest eigenvalue
Dealing with the smallest eigenvalue is less straightforward. Although $\lambda_{\min}(\cdot)$ is a concave function, Jensen's inequality does not help with lower bounding the smallest eigenvalue of the expectation since the matrix $\hat{\Sigma}$ in the argument of this expectation is singular. We therefore take a different route and start by rewriting as follows:

$$\begin{aligned}
& \lambda_{\min}(\Sigma^{\frac{1}{2}}\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)]\Sigma^{\frac{1}{2}}) \\
= \; & \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}(\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R)])^{-1}\Sigma^{-\frac{1}{2}})}
\end{aligned}$$

$$= \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\{\hat{\Sigma} + (\mathrm{E}_R[R^T(R\hat{\Sigma}R^T)^{-1}R])^{-1} - \hat{\Sigma}\}\Sigma^{-\frac{1}{2}})} \tag{4.19}$$

Now, using Weyl's inequality, and the SVD decomposition $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$ combined with Lemma 4.1, the denominator in (4.19) is upper-bounded by:

$$\begin{aligned} & \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{U}\left((\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda}\right)\hat{U}^T\Sigma^{-\frac{1}{2}}) \\ \leqslant\ & \lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda})/\lambda_{\min}(\Sigma) \end{aligned} \tag{4.20}$$

Now observe from lemma 4.2 that the matrix $\mathrm{E}_R[R^T(R\hat{\Lambda}R^T)^{-1}R])^{-1} - \hat{\Lambda}$ is diagonal and, from our analysis in Section 4.1, it has the upper ρ diagonal entries in the interval:

$$\left[(\frac{\rho}{k} - 1)\lambda_{\min\neq 0}(\hat{\Lambda}), (\frac{\rho}{k} - 1)\lambda_{\max}(\hat{\Lambda})\right]$$

and the lower $d - \rho$ diagonal entries in the interval:

$$\left[\frac{\rho - k - 1}{k}\lambda_{\min\neq 0}(\hat{\Lambda}), \frac{\rho - k - 1}{k}\lambda_{\max}(\hat{\Lambda})\right]$$

Hence, $\lambda_{\max}((\mathrm{E}_R[R^T(R\hat{\Lambda}R)^{-1}R])^{-1} - \hat{\Lambda}) \leqslant \frac{\rho}{k}\lambda_{\max}(\hat{\Lambda})$ and so the lower-bounding of (4.20) continues as:

$$\geqslant \frac{1}{\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{4.21}$$

Now observe that $\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}$ is a d-dimensional standard Wishart with $N - 2$ degrees of freedom and scale matrix I_d (e.g. [36] Corr. 3.4.1.2), and using the upper bound in lemma 4.6 for largest eigenvalues of standard Wishart matrices we get (4.21) lower-bounded as

$$\geqslant \frac{1}{(\sqrt{N-2} + \sqrt{d} + \epsilon)^2/N + \frac{\rho}{k}\frac{\lambda_{\max}(\hat{\Lambda})}{\lambda_{\min}(\Sigma)}} \tag{4.22}$$

with probability at least $1 - \exp(-\epsilon^2/2)$.

Finally, we bound $\lambda_{\max}(\hat{\Lambda})$ as:

$$\begin{aligned} \lambda_{\max}(\hat{\Lambda}) &= \lambda_{\max}(\hat{\Sigma}) = \lambda_{\max}(\Sigma\Sigma^{-1}\hat{\Sigma}) \\ &\leqslant \lambda_{\max}(\Sigma)\lambda_{\max}(\Sigma^{-1}\hat{\Sigma}) = \lambda_{\max}(\Sigma)\lambda_{\max}(\Sigma^{-\frac{1}{2}}\hat{\Sigma}\Sigma^{-\frac{1}{2}}) \\ &\leqslant \lambda_{\max}(\Sigma)(\sqrt{N-2} + \sqrt{d} + \epsilon)^2/N \end{aligned}$$

To complete the bound on the condition number we apply union bound and combine the eigenvalue estimates to obtain, after simple algebra, lemma 4.5. □

Back to the proof of Theorem 3.2, substituting into lemma 4.3 the high probability bounds for A and B, rearranging, then setting each of the failure probabilities to $\delta/5$ so that the overall probability of failure remains below δ, then solving for ϵ we obtain Theorem 3.2 after some algebra. For completeness we give these last few straightforward details in Appendix B. □

4.3. Proof of Theorem 3.3

We start from the exact form of the error of FLD in the data space with a fixed training set. Using a similar approach to that employed in proving lemma 4.3, this is easily be shown to be:

$$\begin{aligned}&\Pr(\hat{h}_+(x_q) \neq y_q)\\ &= \sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y} - \hat{\mu}_y)^T \hat{\Sigma}^+ (\hat{\mu}_0 + \hat{\mu}_1 - 2\mu_y)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \hat{\Sigma}^+ \Sigma \hat{\Sigma}^+ (\hat{\mu}_1 - \hat{\mu}_0)}}\right)\end{aligned}$$

where $\hat{\Sigma}^+$ is the pseudo-inverse of the maximum likelihood covariance estimate.
Make the rank ρ SVD decomposition $\hat{\Sigma} = \underline{\hat{U}}\hat{\underline{\Lambda}}\underline{\hat{U}}^T$, where $\underline{\hat{U}}$ is the $d \times \rho$ matrix of eigenvectors associated with the non-zero eigenvalues, $\underline{\hat{U}}^T\underline{\hat{U}} = I_\rho$, and as before $\hat{\underline{\Lambda}}$ is the diagonal $\rho \times \rho$ matrix of non-zero eigenvalues. Then we have:

$$\begin{aligned}&\frac{(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \underline{\hat{U}} \hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \underline{\hat{U}} \hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T \Sigma \underline{\hat{U}} \hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)}}\\ &\leqslant \frac{(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)^T \underline{\hat{U}} \hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)}{\sqrt{\lambda_{\min}(\Sigma)}\sqrt{(\hat{\mu}_1 - \hat{\mu}_0)^T \underline{\hat{U}} \hat{\underline{\Lambda}}^{-2} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)}}\\ &\leqslant \frac{\|\underline{\hat{U}}^T(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)\| \cdot \|\hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}\|\hat{\underline{\Lambda}}^{-1} \underline{\hat{U}}^T (\hat{\mu}_1 - \hat{\mu}_0)\|}\\ &= \frac{\|\underline{\hat{U}}^T(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}}\end{aligned}$$

where we used minorization by Rayleigh quotient and the Cauchy-Schwartz inequality. We will use the well-known fact that $\hat{\Sigma}$ and $\hat{\mu}_1 + \hat{\mu}_0$

are independent [36]. Observe that $\hat{\underline{U}}^T$ is a random matrix with orthonormal rows representing the eigenvectors of the sample covariance of a Gaussian sample. Using the rotational invariance of the multivariate Gaussian distribution, by the Johnson-Lindenstrauss lemma (JLL) this matrix acts as an approximate isometry with high probability [11] that projects a d-dimensional vector onto a random subspace of dimension ρ. Conditioning on $\hat{\mu}_1 + \hat{\mu}_0$ to hold this quantity fixed, and using independence of $\hat{\underline{U}}$ and $\hat{\mu}_1 + \hat{\mu}_0$ [46], we have with probability at least $1 - \exp(-N\epsilon^2/8)$ that:

$$\frac{\|\hat{\underline{U}}^T(\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0)\|}{\sqrt{\lambda_{\min}(\Sigma)}} \leqslant \sqrt{1+\epsilon}\sqrt{\frac{\rho}{d}}\frac{\|\hat{\mu}_1 + \hat{\mu}_0 - 2\mu_0\|}{\sqrt{\lambda_{\min}(\Sigma)}}$$

Further, applying Lemma 4.4 (4.8) to the norm on the r.h.s and replacing in the generalization error expression, we have the following lower bound:

$$\Phi\left(-\frac{1}{2}\sqrt{(1+\epsilon_1)(1+\epsilon_2)}\sqrt{\frac{\rho}{d}\frac{\|\mu_1 - \mu_0\|^2 + \mathrm{Tr}(\Sigma)\frac{N}{N_0N_1}}{\lambda_{\min}(\Sigma)}}\right)$$

with probability at least $1-[\exp(-N\epsilon_1^2/8)+\exp(-\frac{\mathrm{Tr}(\Sigma)+\|\mu_1-\mu_0\|^2\frac{N_0N_1}{N}}{2\lambda_{\max}(\Sigma)}(\sqrt{1+\epsilon_2}-1)^2)]$.

Setting both of these exponential risk probabilities to $\delta/2$ and solving for ϵ_1 and ϵ_2, we obtain the lower bound on the generalization error of pseudoinverted FLD, Theorem 3.3. □

5. Remarks

5.1. On the effect of eigenvector misestimation

We have seen that the eigenvector estimates are not affected by the regularization scheme implemented by our converged ensemble. One may then wonder, since we are dealing with small sample problems, how does misestimation of the eigenvectors of Σ affect the classification performance?

It is known that the quality of eigenvector estimates depends on the eigengaps (differences between ordered eigenvalues) of Σ as well as on the data dimension and number of training examples [49, 32, 41, 48]. Although the sensitivity of eigenvectors to perturbations of matrix entries is well known, the following simple but powerful example from

[31] shows clearly both the problem and the importance of eigenvalue separation. Let:

$$\Sigma = \begin{bmatrix} 1-\epsilon & 0 \\ 0 & 1+\epsilon \end{bmatrix}$$

so that Σ has eigenvalues $1 \pm \epsilon$ and eigenvectors $(1,0)^T$, $(0,1)^T$. On the other hand consider the following perturbed matrix (where the perturbation could arise from, say, estimation error or noise):

$$\Sigma + E = \begin{bmatrix} 1-\epsilon & 0 \\ 0 & 1+\epsilon \end{bmatrix} + \begin{bmatrix} \epsilon & \epsilon \\ \epsilon & -\epsilon \end{bmatrix} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}$$

This matrix also has eigenvalues $1 \pm \epsilon$, but has eigenvectors $\frac{1}{\sqrt{2}}(1,1)^T$, $\frac{1}{\sqrt{2}}(1,-1)^T$, regardless of how small ϵ is.

Applying this in the small sample setting we consider here, if the eigengaps of Σ are too small we can expect bad estimates of its eigenvectors. However, we have seen in Theorem 3.2 that the generalization error of the ensemble can be bounded above by an expression that depends on covariance misestimation only through the condition number of $\hat{S}^{-1}\Sigma \equiv (\Sigma+E)^{-1}\Sigma$ so even a large misestimation of the eigenvectors need not have a large effect on the classification performance: If all the eigengaps are small, so that all the eigenvalues of Σ are similar, then poor estimates of the eigenvectors will not affect this condition number too much. Conversely, following [32] if the eigengaps are large – i.e. we have a very elliptical covariance – then better eigenvector estimates are likely from the same sample size and the condition number of $\hat{S}^{-1}\Sigma$ should still not grow too much as a result of any eigenvector misestimation. In the case of the toy example above, the eigenvalues of $\Sigma(\Sigma+E)^{-1}$ are $\frac{1\pm\epsilon\sqrt{2-\epsilon^2}}{1-\epsilon^2}$, so its condition number is $\frac{1+\epsilon\sqrt{2-\epsilon^2}}{1-\epsilon\sqrt{2-\epsilon^2}}$. For small ϵ this remains fairly close to one – meaning eigenvector misestimation indeed has a negligible effect on classification performance.

5.2. On the effect of k

It is interesting to examine the condition number bound in (4.17) in isolation, and observe the trade off for the projection dimension k which describes very well its role of regularization parameter in the context of our RP-FLD ensemble. To make the numerator smaller k needs to be large while to make the denominator larger it needs to be small. We also see natural behaviour with N, d and the conditioning of the true covariance. From equations (4.13) and (4.16) we see that the condition number bounded by equation (4.17) is the only term in the generalization error bound affected by the choice of k, so we can also partly answer the question left open in [37] about how the optimal k

depends on the problem characteristics, from the perspective of classification performance, by reading off the most influential dependencies that the problem characteristics have on the optimal k. The first term in the numerator of (4.17) contains d but does not contain k while the remaining terms contain k but do not contain d, so we infer that in the setting of $k < \rho - 1 < d$ the optimal choice of k is not affected by the dimensionality d. Noting that for $N < d$ and Gaussian class-conditionals we have $\rho = N - 2$ with probability 1, we see also that for small N or ρ the minimizer of this condition number is achieved by a smaller k (meaning a stronger regulariser), as well as for a small $\kappa(\Sigma)$. Conversely, when N, ρ, or $\kappa(\Sigma)$ is large then k should also be large to minimize the bound.

It is also interesting to note that the regularization scheme implemented by our ensemble has a particularly pleasing form. Shrinkage regularization is the optimal regularizer (w.r.t the Frobenius norm) in the setting when there are sufficient samples to make a full rank estimation of the covariance matrix [34], and therefore one would also expect it to be a good choice for regularization in the range space of $\hat{\Sigma}$. Furthermore ridge regularization in the null space of $\hat{\Sigma}$ can also be considered optimal in the following sense – its effect is to ensure that any query point lying entirely in the null space of $\hat{\Sigma}$ is assigned the maximum likelihood estimate of its class label (i.e. the label of the class with the nearest mean).

5.3. Bias of the Ensemble

By letting $N \to \infty$ (and so $\rho \to d$) while enforcing $k < d = \rho$ we see that our ensemble implements a biased estimate of the true covariance matrix Σ. In particular, plugging in the true parameters μ_y and Σ in the exact error (4.5) we find that the Bayes' risk for FLD in the data space is $\sum_{y=0}^{1} \pi_y \Phi\left(-\frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|\right)$ but the expression in Theorem 3.2 converges to:

$$\sum_{y=0}^{1} \pi_y \Phi\left(-g\left(1 + \frac{d}{k}\kappa(\Sigma)\right)\|\Sigma^{-\frac{1}{2}}(\mu_1 - \mu_0)\|\right)$$

where we recall that $g(1) = \frac{1}{2}$. When $N < d$ however, we see that the generalization error of our RP-FLD ensemble is upper bounded for any training sample containing at least two points for each class whereas our Theorem 3.3 (and asymptotic results in [5]) demonstrate that this is not the case in the data space setting if we regularize by pseudoinverting.

Note that when we plug the expectation examined above into the

classifier ensemble, this is equivalent to an ensemble with infinitely many members and therefore, for any choice of $k < \rho - 1$, although we can underfit (with a poor choice of k) the bounded loss of our ensemble implies that we cannot overfit any worse than the pseudo-inverse FLD data space classifier regardless of the ensemble size, since we do not learn any combination weights from the data. This is quite unlike adaptive ensemble approaches such as AdaBoost, where it is well-known that increasing the ensemble size can indeed lead to overfitting. Furthermore, we shall see from the experiments in the next Section 6 that this guarantee vs. the performance of pseudo-inversion appears to be a conservative prediction of the performance achievable by our randomly-projected ensemble.

5.4. Time Complexities for the RP-FLD Ensemble

We noted in the Introduction that our ensemble, although simple to implement, is also fast. Here we briefly compare the time complexity of our ensemble approach (for a finite ensemble) with that for regularized FLD learnt in the data space.

The time complexity of training a regularized FLD in the data space is dominated by the cost of inverting the estimated covariance matrix $\hat{\Sigma}$ [12], which is $\mathcal{O}(d^3)$ or $\mathcal{O}(d^{\log_2 7}) \simeq \mathcal{O}(d^{2.807})$ using Strassen's algorithm [23].[2] On the other hand, in order to obtain a full-rank inverse covariance estimate in the data space using our ensemble requires $M \in \mathcal{O}(\lceil d/k \rceil)$, and our experimental results in Section 6 suggest that M of this order is indeed enough to get good classification performance. Using this, and taking into account the M $k \times d$ matrix multiplications required to construct the randomly-projected training sets, implies that the time complexity of training our algorithm is $\mathcal{O}(\frac{d}{k}(Nkd + k^3)) = \mathcal{O}(Nd^2 + k^2 d)$ overall, where the k^3 term comes from inverting the full-rank covariance matrix estimate in the projected space. Since we have $k < \rho - 1 < N \ll d$ this is generally considerably faster than learning regularized FLD in the original data space, and furthermore, by using sparse random projection matrices such as those described in [1, 2, 38] one can improve the constant terms hidden by the $\mathcal{O}$ considerably.

For classification on a single core, one can avoid randomly projecting the query point M times by averaging the individual classifiers comprising the ensemble. That is, by bracketing the argument to the ensemble

[2] We note that pseudoinverting $\hat{\Sigma}$ or inverting a diagonal covariance matrix has typical time complexity of $\mathcal{O}(Nd^2)$ or $\mathcal{O}(d)$ respectively. However, as we see from Theorem 3.3 and the experiments in Section 6, the cost in classification performance of these approaches can be prohibitive.

decision rule as:

$$\left((\hat{\mu}_1 - \hat{\mu}_0)^T \frac{1}{M} \sum_{i=1}^{M} R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i\right) \left(x_q - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right)$$

we obtain a single linear classifier of the form $\hat{h} = w + b$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, where:

$$w := \frac{1}{M} \sum_{i=1}^{M} (\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i = \frac{1}{M} \sum_{i=1}^{M} w_i$$

and

$$b := -\frac{1}{M} \sum_{i=1}^{M} (\hat{\mu}_1 - \hat{\mu}_0)^T R_i^T \left(R_i \hat{\Sigma} R_i^T\right)^{-1} R_i \left(\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right) = \frac{1}{M} \sum_{i=1}^{M} b_i$$

Classification of new points using our ensemble then has the same time complexity as classification using the data space FLD, namely $\mathcal{O}(d)$.

6. Experiments

We now present experimental results which show that our ensemble approach is competitive with the state of the art in terms of prediction performance. We do not claim of course that the choice of FLD as a classifier is optimal for these data sets; rather we demonstrate that the various practical advantages of our RP-FLD approach that we listed in the Introduction and Section 5.4, and most importantly its analytical tractability, do not come at a cost in terms of prediction performance.

6.1. Datasets

We used six publicly available high dimensional datasets: Five from the bioinformatics domain (colon, two versions of leukaemia, prostate, and duke breast cancer), and one drug discovery dataset from the 2003 NIPS Feature Selection Challenge (dorothea). The characteristics of these datasets are described in Table I. Our smallest datasets (colon and leukaemia) were the highest dimensional ones used in the empirical RP-classifier study of [20] (although that paper focuses on a single randomly projected classifier vs. the data space equivalent). The 7,129 dimensional leukaemia-large was also the dataset of choice in evaluating a technique for ultrahigh dimensional data in [17]. The 100,000 dimensional dorothea dataset is currently the highest dimensional publicly

Table I. Datasets

Name	Source	#samples	#features
colon	[3]	62	2000
leukaemia	[24]	72	3571
leukaemia large	[24]	72	7129
prostate	[45]	102	6033
duke	[50]	44	7129
dorothea	[26]	800	100000

available dataset in the UCI repository from a problem domain where $N \ll d$ is the norm.

6.2. Protocol

We standardized each data set to have features with mean 0 and variance 1. For dorothea we removed features with zero variance, there were 8402 such features which left a working dimensionality of 91598; we did not do any further feature selection filtering to avoid any external effects in our comparison. For the first five datasets we ran experiments on 100 independent splits, and in each split we took 12 points for testing and used the remainder for training. For dorothea we used the same data split as was used in the NIPS challenge, taking the provided 800 point training set for training and the 350 point validation set for testing. We ran 10 instances for each combination of projection dimension, projection method, and ensemble size - that is 1120 experiments.

For our data space experiments on colon and leukaemia we used FLD with ridge regularization and fitted the regularization parameter using 5-fold cross-validation independently on each training set, following [10], with search in the set $\{2^{-11}, 2^{-10}, ..., 2\}$. However on these data this provided no statistically significant improvement over employing a diagonal covariance in the data space, most likely because of the data scarcity. Therefore for the remaining three bioinformatics datasets (which are even higher dimensional) we used diagonal FLD in the data space. Indeed since diagonal FLD is in use for gene array data sets [13] despite the features being known to be correlated (this constraint acting as a form of regularization) one of the useful benefits of our ensemble is that such a diagonality constraint is no longer necessary.

To satisfy ourselves that building on FLD was a reasonable choice of classifier we also ran experiments in the data space using classical SVM

(using the matlab implementation of [9] on the first five datasets, and the 'liblinear' toolbox [18], which is specialised for very large datasets, for dorothea) and ℓ_1-regularized SVM [18] with linear kernel. In all SVMs the C parameter was fitted by 5-fold cross-validation as above, with search in the set $\{2^{-10}, 2^{-9}, ..., 2^{10}\}$.

For the dorothea dataset it was impractical to consider constructing the full FLD in the dataspace since the covariance matrix would not fit in memory on the authors' machines. We linearized diagonal FLD to get around this issue, but the performance of diagonal FLD was extremely poor (Accuracy of 0.2686) and, since the classical linear SVM is also known to perform poorly on this dataset [28, 27], for the dorothea dataset we baselined against Bernoulli Naïve Bayes (without preprocessing the binary data) following the advice of the challenge organiser to her students given in [28]. We have also run ℓ_1-regularised SVM [18], which turned out successful on this data set.

For all experiments carried out in the projected space, the randomly projected base learners are FLDs with full covariance and no regularization (since we choose $k < \rho - 1$ and so the projected sample covariances are invertible).

6.3. Results

For the five bioinformatics datasets, in each case we compare the performance of the RP ensembles with (regularized) FLD in the data space, vanilla and ℓ_1-regularized SVM, and (as suggested by one of the anonymous referees) with an ensemble of Random Subspace (RS) FLD classifiers[3]. For dorothea we also compare our RP-FLD ensemble with Bernoulli Naïve Bayes.

Summary results for the rule of thumb choice $k = \rho/2$ are listed in Table II as well as end-to-end running times for each data split on a Linux machine with Intel ® Core$^{\text{TM}}$ i5-3570 CPU @ 3.40GHz and 7GB memory for the bioinformatics datasets. The dorothea experiments were run on a Microsoft ® Windows 7$^{\text{TM}}$ machine with the same CPU specification and 8GB memory and, because of the extremely high dimensionality of the dorothea dataset, we refactored our code to avoid randomly projecting the test set by using the approach described in Section 5.4 for these experiments; therefore the running times for

[3] The Random Subspace method [30] consists of projection onto the span of k randomly chosen canonical basis vectors. Note that our theory developed here applies to Gaussian random projection, and this is different to random subspace projection. The RS-FLD decision rule is equivalent to $\hat{h}_P(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)P^T(P\hat{\Sigma}P^T)^+P\left(x_q - \frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)>0\right\}$ where P is a canonical projection matrix and e.g. is therefore not full rank.

dorothea are not directly comparable with those for the biomedical datasets. We note, however, that the main computational overhead for the dorothea dataset comes from the random preprocessing of the data (either random projection, or random subspace) so for these experiments the running times for the preprocessing step are given which still give a good indication of the overall running time.

We see from Table II that with M=1000 members in our ensemble, the SVM outperforms us on two datasets (colon and duke), we outperform it on two datasets (leukaemia-large, and dorothea) and no statistical difference is found on the remaining two datasets. On one dataset (dorothea) ℓ_1-regularised SVM does better than us, we outperform it on three data sets (colon, leukaemia, and leukaemia-large), and there is no statistical difference on the remaining two data sets. We outperform random subspace with 1000 ensemble members on two datasets (duke and prostate) and there is no statistical difference found on the remaining four datasets.

The picture looks not much different for our method having M=100 ensemble members, except there is no significant difference with the ℓ_1-regularised SVM on leukaemia-large, the random subspaces with 100 members beats us on colon and leukaemia, and displays no difference on duke. In fact it turns out that our ensemble with 1000 members differs from that with 100 members on only one data set (duke).

The random subspace FLD ensemble wins over our RP-FLD ensemble with respect to computation time, although this difference is of course confined to the training time only since the time complexity for classification is still $\mathcal{O}(d)$. Interestingly for the random subspace ensembles the overall error performance is just slightly behind that of the random projection ensembles. Since trading off a small amount of accuracy for a speed-up may be desirable for some applications, an interesting research question is whether similar theoretical guarantees to those we obtained for our RP-FLD ensemble can be proved in the random subspace case. Nevertheless the computation time of our RP-FLD ensemble is comparable with the sophisticated liblinear implementation of ℓ_1-regularised SVM, as is its performance. In fact on three of the six data sets tested none of the competing methods outperformed our RP-FLD ensemble at the 95% confidence level.

In figure 1 we plot the results for the regularized data space FLD (Bernoulli Naïve Bayes for dorothea), for a single RP-FLD, and for ensembles of 10, 100, and 3000 RP-FLD classifiers (1000 for dorothea). We see in all cases that our theoretical analysis is well supported, the RP-FLD ensemble outperforms traditional FLD on a range of choices of k and M, and the rule of thumb choice $k = \rho/2$ is not far from the optimal performance – on these data sets $\rho = N - 2$. It is interesting to see that,

Table II. Mean error rates ± 1 standard error, and CPU times estimated from 100 independent splits (10 instances of the fixed split for dorothea) for random projection ensembles with 100 (RP-Ens M=100) or 1000 (RP-Ens M=1000) members, and competing methods (see text for details). For both RP-ensembles and RS-ensembles $k = \rho/2$ was used. The symbols in the t-tests column indicate if the error rates of a competing method is statistically significantly superior (+) or inferior (−) to that of RP-Ensembles in a paired t-test with 95% confidence level. The symbol in the first position is a comparison with RP-Ens with M=100 members and the second symbol is a comparison with RP-Ens with M=1000 members.

Dataset	$\rho/2$	Method	Error	t-tests	CPU Time (sec)
colon	24	RP-Ens M=100	13.50 ± 0.88		0.18 ± 0.002
		RP-Ens M=1000	13.08 ± 0.88		1.48 ± 0.008
		FLD-full	15.50 ± 0.89	−−	10.01 ± 0.064
		SVM	11.58 ± 0.89	++	0.54 ± 0.001
		SVM L1	15.83 ± 1.01	−−	0.53 ± 0.002
		RS-Ens M=100	12.83 ± 0.82	+	0.12 ± 0.025
		RS-Ens M=1000	12.58 ± 0.81		0.88 ± 0.000
leukaemia	29	RP-Ens M=100	2.08 ± 0.40		0.35 ± 0.004
		RP-Ens M=1000	1.67 ± 0.33		3.31 ± 0.029
		FLD-full	2.17 ± 0.39	−	44.99 ± 0.261
		SVM	1.67 ± 0.36		1.09 ± 0.004
		SVM L1	6.08 ± 0.66	−−	1.07 ± 0.003
		RS-Ens M=100	1.83 ± 0.35		0.18 ± 0.000
		RS-Ens M=1000	1.83 ± 0.37		1.81 ± 0.001
leuk-large	29	RP-Ens M=100	2.25 ± 0.44		0.59 ± 0.003
		RP-Ens M=1000	1.92 ± 0.41		6.30 ± 0.056
		FLD-diag	13.33 ± 1.09	−−	0.48 ± 0.003
		SVM	3.50 ± 0.46	−−	2.18 ± 0.012
		SVM L1	2.83 ± 0.55	−	7.03 ± 0.075
		RS-Ens M=100	3.33 ± 0.56	−−	0.44 ± 0.006
		RS-Ens M=1000	2.33 ± 0.49		4.16 ± 0.044
prostate	44	RP-Ens M=100	7.42 ± 0.70		0.82 ± 0.005
		RP-Ens M=1000	7.00 ± 0.70		8.15 ± 0.054
		FLD-diag	38.33 ± 1.57	−−	0.35 ± 0.000
		SVM	7.33 ± 0.72		2.91 ± 0.023
		SVM L1	6.75 ± 0.73		2.85 ± 0.008
		RS-Ens M=100	8.75 ± 0.71	−−	0.56 ± 0.009
		RS-Ens M=1000	8.92 ± 0.73	−−	4.95 ± 0.026
duke	15	RP-Ens M=100	17.50 ± 1.28		0.33 ± 0.002
		RP-Ens M=1000	15.67 ± 1.25	+	3.28 ± 0.023
		FLD-diag	30.58 ± 1.57	−−	0.47 ± 0.000
		SVM	13.50 ± 1.10	++	0.90 ± 0.001
		SVM L1	17.42 ± 1.05		1.14 ± 0.004
		RS-Ens M=100	19.25 ± 1.30	−	0.21 ± 0.000
		RS-Ens M=1000	18.67 ± 1.32	−	2.12 ± 0.002
dorothea	399	RP-Ens M=100	8.66 ± 0.044		211.56 ± 1.944
		RP-Ens M=1000	8.80 ± 0.038		2149.00 ± 24.910
		Bernoulli NB	33.43	−−	4.00
		FLD-diag	71.34	−−	72.98
		SVM	86.86	−−	308.64
		SVM L1	6.00	++	958.53
		RS-Ens M=100	8.57 ± 0.000		122.33 ± 1.312
		RS-Ens M=1000	8.63 ± 0.000		1233.33 ± 10.563

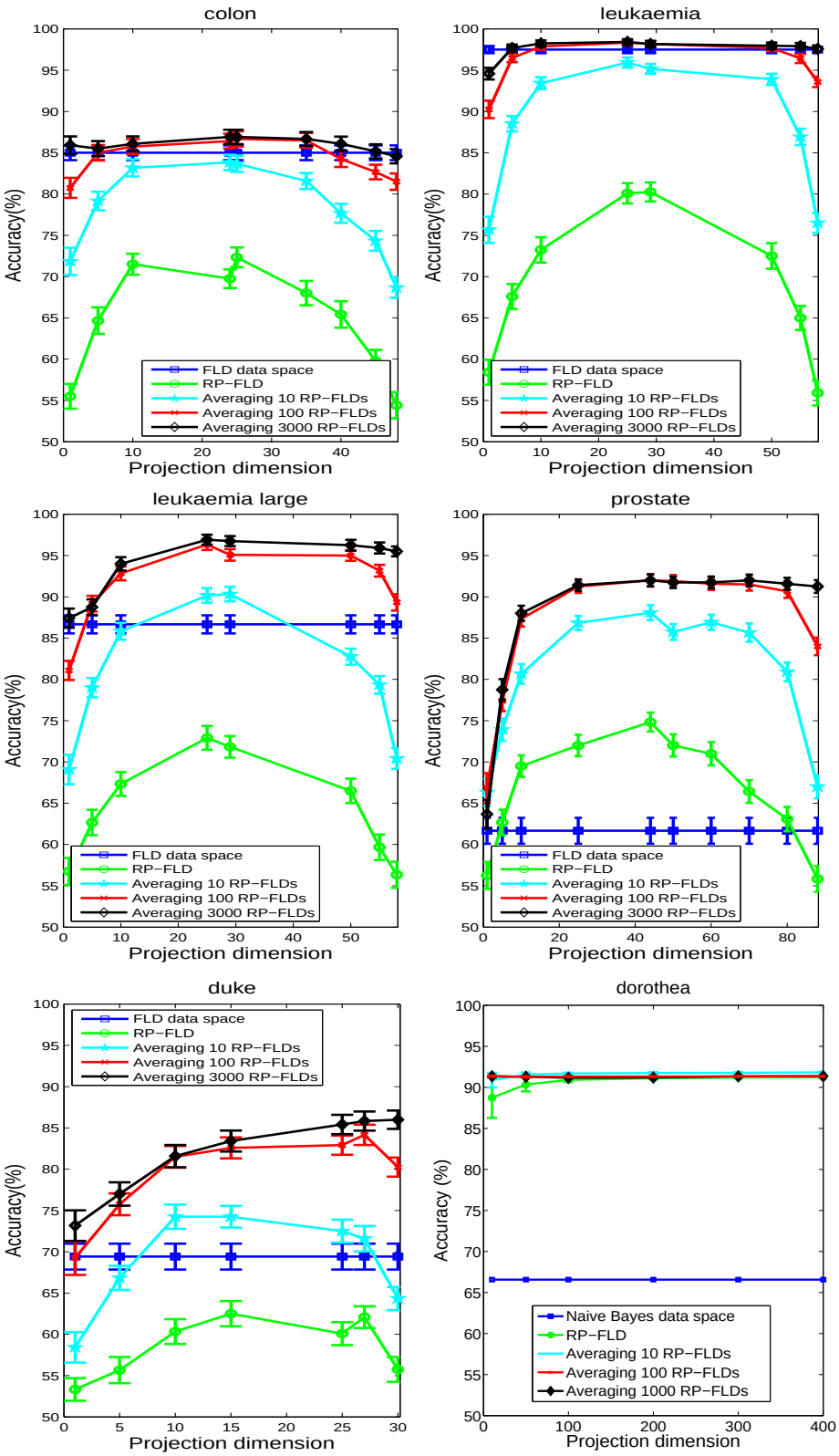

Figure 1. Effect of k. Plots show test error rate versus k and error bars mark 1 standard error estimated from 100 runs (10 repeated runs on the same split for dorothea). In these experiments we used Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries. In each case the projection dimension runs along the x-axis from 1 through to $\rho - 2$.

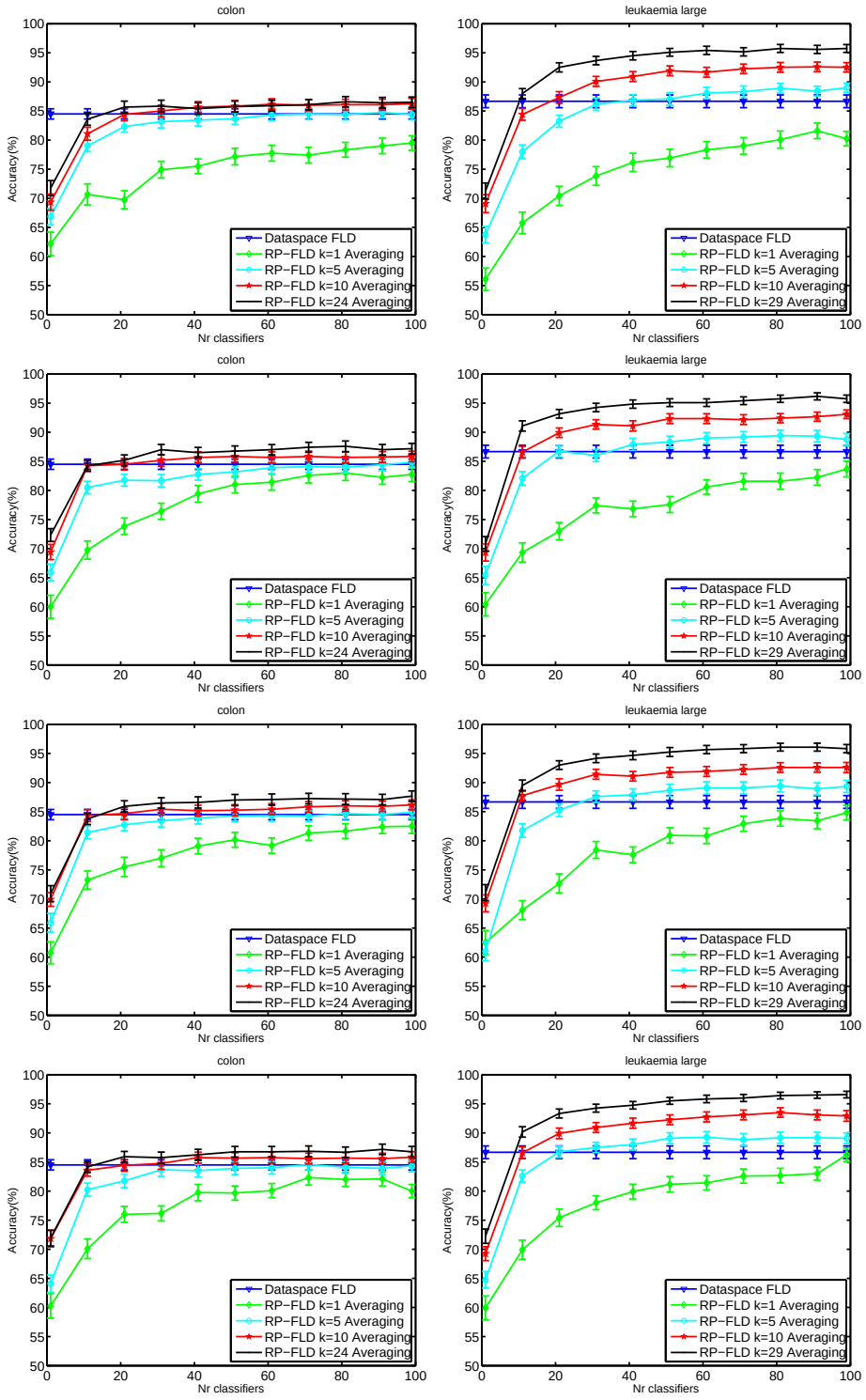

Figure 2. Effect of different random projection matrices and comparison with majority vote. Left hand column shows results on the Colon dataset, right hand column shows results on Leukaemia-large.
Row 1: RP Majority Vote using Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries;
Row 2: RP Averaging using Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries;
Row 3: RP Averaging using ± 1 random matrices with i.i.d entries;
Row 4: RP Averaging using the sparse $\{-1, 0, +1\}$ random matrices from [1].

despite the statistically insignificant difference in performance of full-vs-diagonal covariance models we found for the two lower-dimensional data sets in the data space, for the three higher dimensional data sets (where we used a diagonality constraint for computational tractability) the gap in generalization performance of the data space FLD vs the competing approaches is very large, whereas the gap in performance between the RP-FLD ensembles and the competing approaches is small. Empirically we see, as we might reasonably expect, that capturing the feature covariances via our ensemble approach produces better classification results than working in the data space with a diagonal covariance model.

We ran further experiments on the colon and leukaemia-large data sets to compare the performance of the fast random projections from [1] to Gaussian random projection matrices, and to compare our decision rule to majority vote. Quite interestingly, the picture is very similar and we find no statistically significant difference in the empirical results in comparison with the ensemble that we have presented and analyzed in detail here. The results of these experiments are plotted in figure 2. The performance match between the different choices of random matrix is unsurprising, but the agreement with majority vote is both striking and rather unexpected - we do not yet have an explanation for this behaviour, although it does not appear to arise from the unsigned confidences of the individual ensemble members being concentrated around a particular value.

7. Discussion and Future Work

We considered a randomly projected (RP) ensemble of FLD classifiers and gave theory which, for a fixed training set, explicitly links this ensemble classifier to its data space analogue. We have shown that the RP ensemble implements an implicit regularization of the corresponding FLD classifier in the data space. We demonstrated experimentally that the ensemble can recover or exceed the performance of a carefully-fitted ridge-regularized data space equivalent but with generally lower computational cost. Our theory guarantees that, for most choices of projection dimension k, the error of a large ensemble remains bounded even when the number of training examples is far lower than the number of data dimensions and we gained a good understanding of the effect of our discrete regularization parameter k. In particular, we argued that the regularization parameter k allows us to finesse the known issue of poor eigenvector estimates in this setting. We also demonstrated empirically that with an appropriate choice of k we can obtain good

generalization performance even with few training examples, and a rule of thumb choice $k = \rho/2$ appears to work well.

We showed that, for classification, the optimal choice of k depends on the true data parameters (which are unknown) thereby partly answering – in the negative – the question in [37] concerning whether a simple formula for the optimal k exists.

It would be interesting to extend this work to obtain similar guarantees for ensembles of generic randomly-projected linear classifiers in convex combination, and for an ensemble of random subspace FLDs: we are working on ways to do this. Furthermore, it would be interesting to derive a concentration inequality for matrices in the p.s.d ordering to quantify with what probability a finite ensemble is far from its expectation; this however appears to be far from straightforward – the rank deficiency of $\hat{\Sigma}$ is the main technical issue to tackle.

References

1. Achlioptas, D.: 2003, 'Database-friendly random projections: Johnson-Lindenstrauss with binary coins'. *Journal of Computer and System Sciences* **66**(4), 671–687.
2. Ailon, N. and B. Chazelle: 2006, 'Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform'. In: *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*. pp. 557–563.
3. Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine: 1999, 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays'. *Proceedings of the National Academy of Sciences* **96**(12), 6745.
4. Arriaga, R. and S. Vempala: 1999, 'An algorithmic theory of learning: Robust concepts and random projection'. In: *Foundations of Computer Science, 1999. 40th Annual Symposium on*. pp. 616–623.
5. Bickel, P. and E. Levina: 2004, 'Some theory for Fisher's linear discriminant function, 'naïve Bayes', and some alternatives when there are many more variables than observations'. *Bernoulli* **10**(6), 989–1010.
6. Breiman, L.: 1996, 'Bagging predictors'. *Machine learning* **24**(2), 123–140.
7. Breiman, L.: 2001, 'Random forests'. *Machine learning* **45**(1), 5–32.
8. Brown, G.: 2009, 'Ensemble Learning'. In: C. Sammut and G. Webb (eds.): *Encyclopedia of Machine Learning*. Springer.
9. Cawley, G.C.: 2000, 'MATLAB Support Vector Machine Toolbox (v0.55β) [`http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox`]'. University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ.
10. Cawley, G.C. and N.L. Talbot: 2010, 'On over-fitting in model selection and subsequent selection bias in performance evaluation'. *Journal of Machine Learning Research* **99** 2079–2107.
11. Dasgupta, S. and A. Gupta: 2002, 'An Elementary Proof of the Johnson-Lindenstrauss Lemma'. *Random Structures & Algorithms* **22**, 60–65.
12. Duda, R., P. Hart, and D. Stork: 2000, *Pattern Classification*. Wiley, 2 edition.

13. Dudoit, S., J. Fridlyand, and T. Speed: 2002, 'Comparison of discrimination methods for the classification of tumors using gene expression data'. *Journal of the American statistical association* **97**(457), 77–87.
14. Durrant, R. and A. Kabán: 2010, 'Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data'. In: *Proceedings16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*.
15. Durrant, R. and A. Kabán: 2012, 'Error bounds for Kernel Fisher Linear Discriminant in Gaussian Hilbert space'. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AIStats 2012)*.
16. Durrant, R. J.: 2013, 'Learning in High Dimensions with Projected Linear Discriminants'. Ph.D. thesis, School of Computer Science, University of Birmingham.
17. Fan, J. and J. Lv: 2008, 'Sure independence screening for ultrahigh dimensional feature space'. *J. R. Statist. Soc. B* **70**, Part 5, 849–911.
18. Fan, R., K. Chang, C. Hsieh, X. Wang, and C. Lin: 2008, 'LIBLINEAR: A Library for Large Linear Classification'. *Journal of Machine Learning Research* **9**, 1871–1874.
19. Folgieri, R.: 2008, 'Ensembles based on Random Projection for gene expression data analysis'. Ph.D. thesis, Università degli Studi di Milano.
20. Fradkin, D. and D. Madigan: 2003, 'Experiments with random projections for machine learning'. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 522–529.
21. Fumera, G., F. Roli, and A. Serrau: 2008, 'A theoretical analysis of bagging as a linear combination of classifiers'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **30**(7), 1293–1299.
22. Goel, N., G. Bebis, and A. Nefian: 2005, 'Face recognition experiments with random projection'. In: *Proceedings of SPIE*, Vol. 5779. p. 426.
23. Golub, G. and C. Van Loan: 2012, *Matrix computations*, Vol. 3. JHU Press.
24. Golub, T., D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander: 1999, 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring'. *Science* **286**(5439), 531.
25. Guo, Y., T. Hastie, and R. Tibshirani: 2007, 'Regularized linear discriminant analysis and its application in microarrays'. *Biostatistics* **8**(1), 86–100.
26. Guyon, I., NIPS 2003 Feature Selection Challenge: Dorothea dataset [`http://www.nipsfsc.ecs.soton.ac.uk/datasets/DOROTHEA.zip`]. Retrieved from internet 14th April 2014.
27. Guyon, I., S.R. Gunn, A. Ben-Hur, and G. Dror: 2004 'Result Analysis of the NIPS 2003 Feature Selection Challenge'. *NIPS* **4**, 545–552.
28. Guyon, I., J. Li, T. Mader, P. Pletscher, G. Schneider, and M. Uhr: 2006, 'Feature Selection with the CLOP Package'. Technical report, [`http://clopinet.com/isabelle/Projects/ETH/TM-fextract-class.pdf`]. Retrieved from internet 14th April 2014.
29. Hastie, T., R. Tibshirani, and J. Friedman: 2001, *The elements of statistical learning; data mining, inference, and prediction*. Springer.
30. Ho, T.: 1998, 'The random subspace method for constructing decision forests'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **20**(8), 832–844.
31. Horn, R. and C. Johnson: 1985, *Matrix Analysis*. Cambridge University Press.

32. Johnstone, I. and A. Lu: 2009, 'On consistency and sparsity for principal components analysis in high dimensions'. *Journal of the American Statistical Association* **104**(486), 682–693.
33. Koltchinskii, V. and D. Panchenko: 2002, 'Empirical margin distributions and bounding the generalization error of combined classifiers'. *The Annals of Statistics* **30**(1), 1–50.
34. Ledoit, O. and M. Wolf: 2004, 'A well-conditioned estimator for large-dimensional covariance matrices'. *Journal of multivariate analysis* **88**(2), 365–411.
35. Maniglia, S. and A. Rhandi: 2004, 'Gaussian measures on separable Hilbert spaces and applications'. *Quaderni del Dipartimento di Matematica dell'Università del Salento* **1**(1), 1–24.
36. Mardia, K., J. Kent, and J. Bibby: 1979, *Multivariate analysis*. London: Academic Press.
37. Marzetta, T., G. Tucci, and S. Simon: 2011, 'A Random Matrix–Theoretic Approach to Handling Singular Covariance Estimates'. *IEEE Trans. Information Theory* **57**(9), 6256–71.
38. Matoušek, J.: 2008, 'On variants of the Johnson–Lindenstrauss lemma'. *Random Structures & Algorithms* **33**(2), 142–156.
39. Mika, S., G. Ratsch, J. Weston, B. Schölkopf, and K. Mullers: 2002, 'Fisher discriminant analysis with kernels'. In: *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*. pp. 41–48.
40. Pattison, T. and D. Gossink: 1999, 'Misclassification Probability Bounds for Multivariate Gaussian Classes'. *Digital Signal Processing* **9**, 280–296.
41. Paul, D. and I. Johnstone: 2012, 'Augmented sparse principal component analysis for high dimensional data'. *arXiv preprint arXiv:1202.1242*.
42. Raudys, S. and R. Duin: 1998, 'Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix'. *Pattern Recognition Letters* **19**(5), 385–392.
43. Schapire, R., Y. Freund, P. Bartlett, and W. Lee: 1998, 'Boosting the margin: A new explanation for the effectiveness of voting methods'. *The Annals of Statistics* **26**(5), 1651–1686.
44. Schclar, A. and L. Rokach: 2009, 'Random Projection Ensemble Classifiers'. In: J. Filipe, J. Cordeiro, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski (eds.): *Enterprise Information Systems*, Vol. 24 of *Lecture Notes in Business Information Processing*. Springer, pp. 309–316.
45. Singh, D., P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers: 2002, 'Gene expression correlates of clinical prostate cancer behavior'. *Cancer cell* **1**(2), 203–209.
46. Tulino, A. and S. Verdú: 2004, *Random matrix theory and wireless communications*. Now Publishers Inc.
47. Vershynin, R.: 2012, 'Introduction to Non-asymptotic Random Matrix Theory'. In: Y. Eldar and G. Kutyniok (eds.): *Compressed Sensing, Theory and Applications*. Cambridge University Press, pp. 210–268.
48. Vu, V.: 2011, 'Singular vectors under random perturbation'. *Random Structures & Algorithms* **39**(4), 526–538.
49. Vu, V. and J. Lei: 2012, 'Minimax Rates of Estimation for Sparse PCA in High Dimensions'. In: *Proceedings 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, Vol. 22. pp. 1278–1286.

50. West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins: 2001, 'Predicting the clinical status of human breast cancer by using gene expression profiles'. *Proceedings of the National Academy of Sciences* **98**(20), 11462.

Appendix

A. Proof of Lemma 4.4

We prove the statement of eq. (4.8) fully, and outline the proof of (4.9) which is very similar. Let $t > 0$ be a positive real constant (to be optimized later), then:

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\} \\
&= \Pr\left\{\exp\left(t\|X\|^2\right) \geqslant \exp\left(t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\} \\
&\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\mathrm{E}\left[\exp\left(t\|X\|^2\right)\right] \qquad \text{(A.1)}
\end{aligned}$$

Where (A.1) follows by Markov's inequality. Now, $X \sim \mathcal{N}(\mu, \Sigma)$ and so $\|X\|^2 = \sum_{i=1}^{d} X_i^2$ has a non-central χ^2 distribution, and therefore $\mathrm{E}\left[\exp\left(t\|X\|^2\right)\right]$ is the moment generating function of a non-central χ^2 distribution. Hence (e.g. [35] proposition 1.2.8) for all $t \in (0, 1/2\lambda_{\max}(\Sigma))$ we have (A.1) is equal to:

$$\begin{aligned}
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{d}\left(1-2t\lambda_i\right)^{-\frac{1}{2}}\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_i}\right) \\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{d}\left(1+\tfrac{2t\lambda_i}{1-2t\lambda_i}\right)^{\frac{1}{2}}\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_i}\right) \\
&\leqslant \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\prod_{i=1}^{d}\exp\left(\tfrac{1}{2}\tfrac{2t\lambda_i}{1-2t\lambda_{\max}(\Sigma)}\right)\exp\left(\tfrac{t\mu_i^2}{1-2t\lambda_{\max}(\Sigma)}\right) \\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\tfrac{t\left(\sum_{i=1}^{d}\lambda_i+\mu_i^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right) \\
&= \exp\left(-t\left(1+\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)+\tfrac{t\left(\mathrm{Tr}(\Sigma)+\|\mu\|^2\right)}{1-2t\lambda_{\max}(\Sigma)}\right) \qquad \text{(A.2)}
\end{aligned}$$

Now taking $t = \frac{1-(1+\epsilon)^{-\frac{1}{2}}}{2\lambda_{\max}(\Sigma)} \in (0, 1/2\lambda_{\max}(\Sigma))$ and substituting this value of t into (A.2) yields, after some algebra, (4.8):

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \geqslant (1+\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\} \\
&\leqslant \exp\left(-\tfrac{\mathrm{Tr}(\Sigma)+\|\mu\|^2}{2\lambda_{\max}(\Sigma)}\left(\sqrt{1+\epsilon}-1\right)^2\right)
\end{aligned}$$

The second inequality (4.9) is proved similarly. We begin by noting:

$$\begin{aligned}
&\Pr\left\{\|X\|^2 \leqslant (1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right\} \\
&= \Pr\left\{\exp\left(-t\|X\|^2\right) \geqslant \exp\left(-t\left(1-\epsilon\right)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)\right)\right\} \\
&\leqslant \exp\left(t(1-\epsilon)\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)-t\left(\mathrm{Tr}\left(\Sigma\right)+\|\mu\|^2\right)/1+2t\lambda_{\max}(\Sigma)\right)
\end{aligned}$$

and then complete the proof as before, substituting in the optimal $t = \frac{1+(1-\epsilon)^{-\frac{1}{2}}}{2\lambda_{\max}(\Sigma)}$ to give the bound.

B. Details for the end of proof of Theorem 3.2

There are five terms to simultaneously bound with high probability, namely the two B_y, A, and the two extreme eigenvalues involved in the condition number bound. We use the standard approach of setting each of the confidence probabilities no greater than $\delta/5$ and solving for ϵ (or a function of ϵ appearing in the bound) then back-substituting and applying the union bound to derive a guarantee which holds with probability $1-\delta$.
Firstly, for the extreme eigenvalues we have (twice):

$$\begin{aligned} \exp\left(-\epsilon_3^2/2\right) &\leqslant \delta/5 \\ \implies \sqrt{2\log(5/\delta)} &\leqslant \epsilon_3 \end{aligned} \tag{B.1}$$

For the upper bounds on the B_y we have:

$$\exp\left(-\frac{d}{2}\left(\sqrt{1+\epsilon_y}-1\right)^2\right) \leqslant \delta/5$$

and solving for $\sqrt{1+\epsilon_y}$ we obtain:

$$\begin{aligned} \sqrt{\frac{2\log(5/\delta)}{d}} &\leqslant \pm\left(\sqrt{1+\epsilon_y}-1\right) \\ \implies 1+\sqrt{\frac{2\log(5/\delta)}{d}} &\geqslant \sqrt{1+\epsilon_y} \end{aligned} \tag{B.2}$$

Finally, for the lower bound on A (which holds for both classes simultaneously) we solve for $\sqrt{1-\epsilon_2}$ to obtain:

$$\begin{aligned} &\exp\left(-\left(\frac{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}{2N/N_0N_1}\right)\left(\sqrt{1-\epsilon_2}-1\right)^2\right) \leqslant \delta/5 \\ \iff &\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2} \leqslant \left(\sqrt{1-\epsilon_2}-1\right)^2 \\ \iff &\sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} \leqslant \pm\left(\sqrt{1-\epsilon_2}-1\right) \\ \implies &1-\sqrt{\frac{2N\log(5/\delta)/N_0N_1}{dN/N_0N_1+\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2}} \geqslant \sqrt{1-\epsilon_2} \end{aligned} \tag{B.3}$$

Plugging the left hand sides of the inequalities (B.1), (B.2) and (B.3) into the bounds on κ, B_0, B_1 and A for ϵ_3, $\sqrt{1+\epsilon_0}$, $\sqrt{1+\epsilon_1}$ and $\sqrt{1-\epsilon_2}$ respectively gives, after some algebra, the stated Theorem 3.2.