JMLR: Workshop and Conference Proceedings 29:1–16, 2013 ACML 2013

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

Robert J. Durrant BOBD@WAIKATO.AC.NZ
Department of Statistics, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand.

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

Editor: Cheng Soon Ong and Tu Bao Ho

Abstract

We examine the performance of an ensemble of randomly-projected Fisher Linear Discriminant classifiers, focusing on the case when there are fewer training observations than data dimensions. Our ensemble is learned from a sequence 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.

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.

We confirm theoretical findings with experiments, and demonstrate the utility of our approach on several datasets from the bioinformatics domain where fewer observations than dimensions are the norm.

Keywords: Random Projections, Linear Discriminant Analysis, Ensemble Classifiers.

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 (Breiman, 1996); random subspaces (Ho, 1998); random forests (Breiman, 2001). Surprisingly, despite the well-known theoretical properties of random projections as dimension-reducing approximate isometries (Dasgupta and Gupta, 2002; Achlioptas, 2003) and empirical and theoretical studies demonstrating their usefulness when learning a *single* classifier (e.g. Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Durrant and Kabán, 2010), 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 (Goel et al., 2005; Folgieri, 2008; Schclar and Rokach, 2009) all report good 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 Fumera et al. (2008) 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 (e.g. Schapire et al., 1998), such analysis only serves to relate the performance of an ensemble to its members and Fumera et al. (2008) 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 Marzetta et al. (2011) 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: 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, parallel implementation of our approach is straightforward since, both for training and classification, the individual ensemble members can be run on separate cores and the ensemble decision is then given by simply summing the individual classifier outputs. 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 which are 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 $N' \leqslant N$ training examples uniformly with replacement from a training set of size 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 Schapire et al. (1998), 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 (Guo et al., 2007; Dudoit et al., 2002). Our experiments suggest that in practice, when the number of training examples is less than the number of data dimensions, this 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 $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$ where $x_i \overset{i.i.d}{\sim} \mathcal{D}_x$ and each x_i has an associated label $y_i \in \{0, 1\}$. We take the $x_i \in \mathbb{R}^d$ to be d-dimensional real valued observations. 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 $N < d$. 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 (Raudys and Duin, 1998).
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 $r_{ij} \sim \mathcal{N}(0, \sigma^2)$. We can take $\sigma^2 = 1$ without loss of generality. Such matrices are called random projection matrices in the literature (Arriaga and Vempala, 1999; Achlioptas, 2003)[1].

1. We find empirically that, as one would expect, other common choices of random projection matrix with zero-mean i.i.d sub-Gaussian entries (e.g. Achlioptas, 2003) do not affect the ensemble performance.

Pre-multiplying the data with one of the matrices R maps the training examples to a k-dimensional subspace of the data space $\mathbb{R}^d$ and, 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 (Brown, 2009). 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:

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

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 begin 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 2 we consider random instantiations of the training set in order to give a tail bound on the generalization error of the ensemble.

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

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

provided that this limit exists. 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 $k \in \{1,...,\rho-2\} \cup \{\rho+2,...,d\}$, then this expectation is indeed defined for each entry. From equation (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 (with respect to random matrices R):

$$\mathrm{E}\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right] \qquad (2)$$

Before continuing, we should note that for the case $k \in \{1,...,\rho-2\}$ Marzetta et al. (2011) 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 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 Marzetta et al. (2011), 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. Results

Our main theoretical results are the following two 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.

Theorem 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 $\hat{S}^{-1} := E\left[R^T \left(R\hat{\Sigma}R^T\right)^{-1} R\right]$, and denote by $\kappa(\hat{S}^{-1})$ its condition number, $\kappa(\hat{S}^{-1}) = \lambda_{\max}(\hat{S}^{-1})/\lambda_{\min}(\hat{S}^{-1})$. Then:*

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

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,

$$\hat{h}_R(x_q) := \mathbf{1}\left\{(\hat{\mu}_1 - \hat{\mu}_0)R^T(R\hat{\Sigma}R^T)^{-1}R\left(x_q - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right) > 0\right\} \tag{3}$$

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) 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 2, (an upper bound on) the generalization error of FLD increases as a function of this condition number. In turn, the bound given in our Theorem 1 depends on the extreme *non-zero* eigenvalues of $\hat{\Sigma}$, its rank[2] 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 2 (Generalization error of the converged ensemble) *Let $\mathcal{T}_N = \{(x_i, y_i)\}_{i=1}^N$ 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 $x_q|y_q \sim \mathcal{N}(\mu_y, \Sigma)$, 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,*

2. In the setting considered here we typically have $\rho = N - 2$

the generalization error of the converged ensemble of randomly projected FLD classifiers is upper-bounded with probability at least $1-\delta$ *by the following:*

$$\Pr_{x_q,y_q}(\hat{h}_{ens}(x_q)\neq y_q) \leqslant \sum_{y=0}^{1}\pi_y\Phi\left(-\left[g\left(\bar{\kappa}\left(\sqrt{2\log\frac{5}{\delta}}\right)\right)\times\ldots\right.\right. \tag{4}$$
$$\left.\left.\ldots\left[\sqrt{\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2+\frac{dN}{N_0N_1}}-\sqrt{\frac{2N}{N_0N_1}\log\frac{5}{\delta}}\right]_+-\sqrt{\frac{d}{N_y}}\left(1+\sqrt{\frac{2}{d}\log\frac{5}{\delta}}\right)\right]\right)$$

where $\bar{\kappa}(\epsilon)$ *is a high probability (w.r.t draws of* $\mathcal{T}_N$*) upper bound on the condition number of* $\Sigma\hat{S}^{-1}$ *given by eq. (12) 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 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.

4. Proofs

4.1. Proof of Theorem 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 give a full proof for the case $k=1$ for the sake of argument and insight – details for all other cases of $k<d$ can be found in the supplementary material. To smooth our way we will make use of the following two easy, but useful, lemmas:

Lemma 3 (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 (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* $\hat{U}\hat{\Lambda}\hat{U}^T$, *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 Marzetta et al. (2011).

It follows from lemmas 3 and 4 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. For reasons of space here we only give a full proof for the case $k = 1$, when we are projecting the high dimensional data on to a single line for each classifier in the ensemble. The other cases are dealt with in the supplementary material, as previously noted. In the case $k = 1$ 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 3 and 4 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:

$$\kappa\left(\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]\right) \leqslant \frac{\rho}{\rho-2}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{5}$$

whereas in the setting $N < d$ the ML covariance estimate has unbounded condition number. To briefly sketch out our approach 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:

$$\left[\begin{array}{c|c} \mathrm{E}[P^T\left(P\hat{\underline{\Lambda}}P^T\right)^{-1}P] & 0 \\ \hline 0 & \mathrm{E}[S^T\left(P\hat{\underline{\Lambda}}P^T\right)^{-1}S]\end{array}\right] \tag{6}$$

Here in $P\hat{\underline{\Lambda}}P^T$ we use $\hat{\underline{\Lambda}}$ to denote the $\rho \times \rho$ positive definite upper block of the positive semi-definite matrix $\hat{\Lambda}$. We then show that the diagonal elements in the upper block are all in the interval:

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

Hence, in the upper block the condition number is reduced compared to that of $\hat{\Lambda}$ restricted to its range.

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

In other words, in the range of $\hat{\Sigma}$, the ensemble has the effect of a shrinkage regularizer (Ledoit and Wolf, 2004). Similarly we show that the lower block is a multiple of $I_{d-\rho}$ with the coefficient in the interval:

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

That is, in the null space of $\hat{\Sigma}$ the ensemble acts as a ridge regularizer (Hastie et al., 2001).

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

$$\kappa\left(\mathrm{E}\left[R^T\left(R\hat{\Lambda}R^T\right)^{-1}R\right]\right) \leqslant \frac{\rho}{\rho-k-1}\cdot\frac{\lambda_{\max}}{\lambda_{\min\neq 0}} \tag{7}$$

which we see reduces to eq.(5) when $k=1$.

4.2. Sketch Proof of Theorem 2

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 with a fixed training set

Traditionally, ensemble methods are regarded as 'meta-learning' approaches and although bounds exist (e.g. Koltchinskii and Panchenko, 2002) 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 an FLD ensemble 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. We note that an upper bound on generalization error with similar behaviour can be derived for the much larger class of sub-Gaussian distributions with different covariance matrices (see e.g. Durrant and Kabán, 2010), therefore this Gaussianity assumption is not crucial.

Lemma 5 (Exact generalization error with Gaussian classes) *Let $x_q|y_q \sim \mathcal{N}(\mu_y, \Sigma)$ and let $Pr\{y_q = y\} = \pi_y$, where $\Sigma \in \mathcal{M}_{d\times d}$ is a full rank covariance matrix. Let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix with i.i.d. 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 (1) is given by:*

$$\Pr_{(x_q,y_q)}\{\hat{h}_{ens}(x_q) \neq y_q\} = \sum_{y=0}^{1}\pi_y\Phi\left(-\frac{1}{2}\frac{(\hat{\mu}_{\neg y}-\hat{\mu}_y)^T\hat{S}^{-1}(\hat{\mu}_0+\hat{\mu}_1-2\mu_y)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}}\right) \tag{8}$$

The proof of this lemma is similar in spirit to the one for a single FLD given in Bickel and Levina (2004); Pattison and Gossink (1999), and we omit it here. We note however that equation (8) has the same form as the error of the data space FLD (*Ibid.*) 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. Proof (sketch) of the tail bound on the generalization error

Now we briefly sketch out how we can bound the generalization error of the RP-FLD ensemble with high probability over the random draw of a training set of size $N = N_0 + N_1$. We begin by decomposing the numerator of the generalization error term (for a single class) obtained in Lemma 5 as follows:

$$\begin{aligned}&(\hat{\mu}_1+\hat{\mu}_0-2\mu_0)^T\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\\ =\;&(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)+2(\hat{\mu}_0-\mu_0)^T\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)\end{aligned} \tag{9}$$

Using this decomposition we can rewrite the argument of the first term in Lemma 5 in the following form:

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

Where:

$$A = \frac{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}} \tag{10}$$

and:

$$B = \frac{2(\mu_0-\hat{\mu}_0)^T\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}{\sqrt{(\hat{\mu}_1-\hat{\mu}_0)^T\hat{S}^{-1}\Sigma\hat{S}^{-1}(\hat{\mu}_1-\hat{\mu}_0)}} \tag{11}$$

We lower bound A and upper bound B to bound the whole term from above with high probability and, since Φ is monotonic increasing in its argument, this yields an upper bound on generalization error. Omitting the details here, we show in the supplementary material that the following are true:

1. Define:
$$\bar{\kappa}(\epsilon) := \frac{(\sqrt{N-2}+\sqrt{d}+\epsilon)^2(1+\rho/k\cdot\kappa(\Sigma))}{(\sqrt{N-2}-\sqrt{k}-\epsilon)^2} \tag{12}$$
then A is bounded below by:
$$A \geqslant 2g(\bar{\kappa}(\epsilon))\sqrt{(1-\epsilon)\left(\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2+\frac{d\cdot N}{N_0N_1}\right)} \tag{13}$$
with probability at least:
$$1-\exp\left(-\left(\frac{d}{2}+\frac{\|\Sigma^{-\frac{1}{2}}(\mu_1-\mu_0)\|^2N_0N_1}{2N}\right)\left(\sqrt{1-\epsilon}-1\right)^2\right)-2\exp(-\epsilon^2/2) \tag{14}$$

2. B is bounded above by:
$$B \leqslant 2\sqrt{(1+\epsilon)d/N_0} \tag{15}$$
with probability at least $1-\exp(-\frac{d}{2}(\sqrt{1+\epsilon}-1)^2)$.

Substituting in Theorem 2 these bounds for A and B, rearranging, then setting each of the failure probabilities no greater than $\delta/5$ so that the overall probability of failure remains below δ, and solving for ϵ we obtain the theorem after some algebra.

4.3. Comments

1. Observe that the strength of the regularization depends on k and ρ, and the non-zero eigenvalues of $\hat{\Sigma}$. More precisely, $\frac{k}{\rho-k-1}$ and $\frac{k}{\rho}$ increase monotonically with k (and decrease with ρ). Since we are talking about an inverse covariance estimate, this implies that the extent of regularization decreases with increasing k (and increases when ρ gets larger). Hence, k takes the role of the regularization parameter and the analysis in this and the following subsections provides us with insight for setting this parameter. For the data sets we used in our experiments $k \simeq \rho/2$ appears to be a reasonable rule of thumb choice.

2. The regularization scheme implemented by our ensemble has a particularly pleasing form. Shrinkage regularization is the optimal regularizer (with respect to the Frobenius norm) in the setting when there are sufficient samples to make a full rank estimation of the covariance matrix (Ledoit and Wolf, 2004), and therefore one would also expect it to be a good choice for regularization in the range 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).

3. We can show that letting $N \to \infty$ (and so $\rho \to d$) while enforcing $k < d = \rho$ 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 (8) 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 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 one point for each class whereas it is known (Bickel and Levina, 2004) that this is not the case in the data space setting if we regularize by pseudo-inverting.

4. As $k \nearrow \rho - 1$ we can show that the values in the upper block approach the non-zero eigenvalues of $\hat{\Sigma}^+$ while in the lower block the diagonal entries become extremely large, and when $k = \rho - 1$ we recover precisely the data space pseudo-inverse performance. Hence when $k \simeq \rho$ we overfit about as badly as pseudo-inverting in the data space. 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) we cannot overfit any worse than the unregularized (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 5 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. 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 the RP-FLD approach we listed in the Introduction do not come at a cost in terms of prediction performance.

5.1. Datasets

We used five publicly available high dimensional datasets from the bioinformatics domain (colon, two versions of leukemia, prostate, and duke breast cancer), whose characteristics are as described in Table 1. The first two (colon and leukemia) have the smallest dimensionality amongst these and were the highest dimensional data sets used in the empirical RP-classifier study of Fradkin and Madigan (2003) (Note, that paper focuses on a single randomly projected classifier vs. the data space equivalent, and does not consider RP-FLD).

5.2. Protocol

We standardized each data set to have features with mean 0 and variance 1, and ran experiments on 100 independent splits. In each split we took 12 points for testing and

Table 1: Datasets

Name	Source	#samples	#features
colon	Alon et al. (1999)	62	2000
leukemia	Golub et al. (1999)	72	3571
leukemia large	Golub et al. (1999)	72	7129
prostate	Singh et al. (2002)	102	6033
duke	West et al. (2001)	44	7129

Table 2: Mean error rates $\pm$ 1 standard error, estimated from 100 independent splits when $k = \rho/2$.

Dataset	$\rho/2$	100 RP-FLD	1000 RP-FLD	SVM
colon	24	13.58 ± 0.89	13.08 ± 0.86	16.58 ± 0.95
leuk.	29	1.83 ± 0.36	1.83 ± 0.37	1.67 ± 0.36
leuk.lge	29	4.91 ± 0.70	3.25 ± 0.60	3.50 ± 0.46
prost.	44	8.00 ± 0.76	8.00 ± 0.72	8.00 ± 0.72
duke	15	17.41 ± 1.27	16.58 ± 1.27	13.50 ± 1.10

used the remainder for training. For our data space experiments on colon and leukemia we used FLD with ridge regularization and fitted the regularization parameter using 5-fold cross-validation on the first five data splits following Mika et al. (2002). 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 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 (Dudoit et al., 2002) 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.

The randomly projected base learners are instances of FLD with full covariance learned in the projected space. To satisfy ourselves that building on FLD was a reasonable choice of classifier we also ran experiments in the data space using SVM with typical default settings[3], as was done in Fradkin and Madigan (2003).

5.3. Results

In each case we compare the performance of our RP averaging ensemble with (regularized) FLD in the data space and also with SVM in the data space. Summary results for the rule of thumb choice $k = \rho/2$ are listed in Table 2.

3. MATLAB support vector machine toolbox (Cawley, 2000) with linear kernel and the parameter C set to $C = 1$.

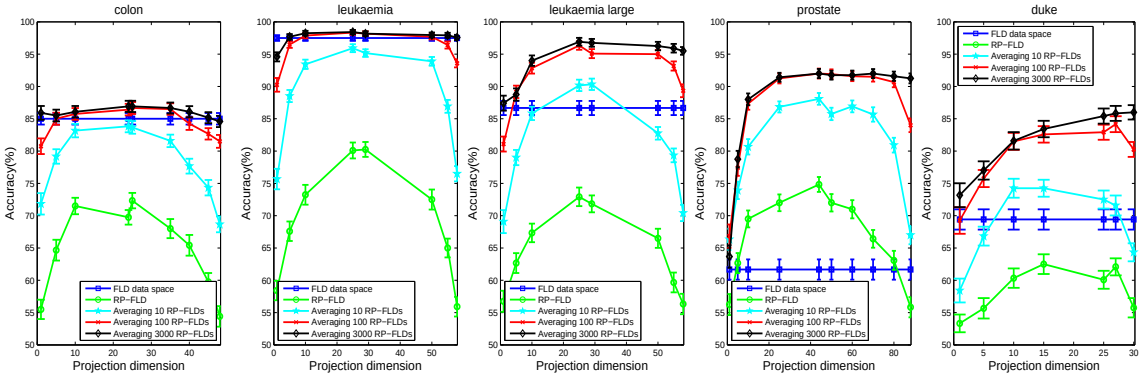

Figure 1: Effect of k. Plots show test error rate versus k and error bars mark 1 standard error estimated from 100 runs. In these experiments we used Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries.

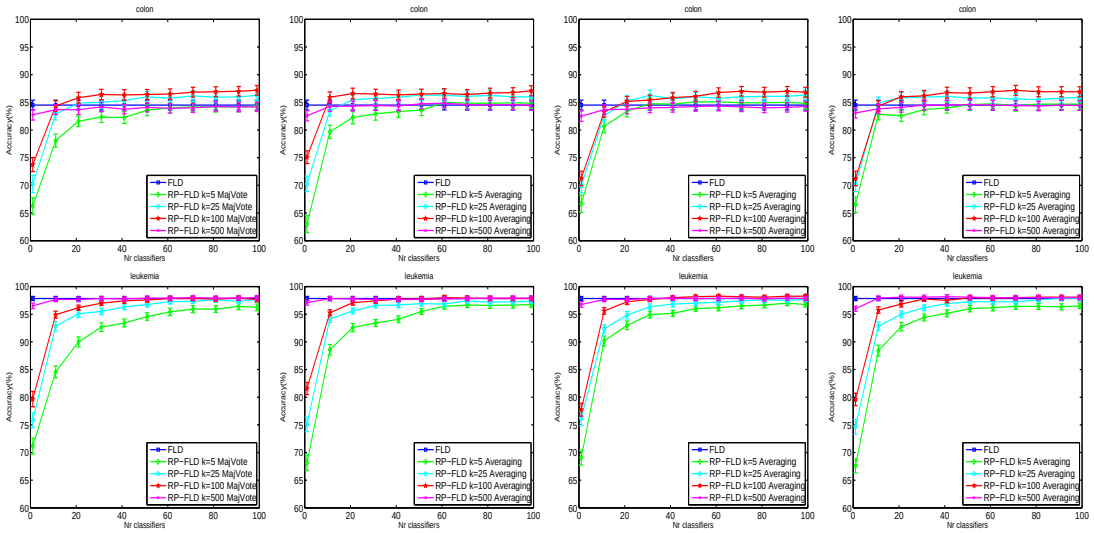

Figure 2: Effect of different random projection matrices and comparison with majority vote. Column 1: RP Majority Vote using Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries; Column 2: RP Averaging ensemble using Gaussian random matrices with i.i.d $\mathcal{N}(0,1)$ entries; Column 3: RP Averaging ensemble using ± 1 random matrices with i.i.d entries; Column 4: RP Averaging ensemble using the sparse $\{-1, 0, +1\}$ random matrices from Achlioptas (2003).

In figure 1 we plot the results for the regularized data space FLD, for a single RP-FLD, and for ensembles of 10, 100, and 3000 RP-FLD classifiers. 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 the rule of thumb choice $k = \rho/2$ is not far from the optimal performance. It is interesting to see that, 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 SVM is very large, whereas the gap in performance between the RP-FLD ensembles and SVM 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 leukemia data sets to compare the performance of the fast random projections from Achlioptas (2003) 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.

6. 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. We also demonstrated empirically that we can obtain good generalization performance even with few training examples, and a rule of thumb choice $k = \rho/2$ appears to work well. It would be interesting to extend this work to obtain similar guarantees for ensembles of generic randomly-projected linear classifiers in convex combination, and we are working on ways to do this. Furthermore, it would be interesting to obtain high probability guarantees on the performance of a finite ensemble, e.g. by deriving a concentration inequality for matrices in the positive semi-definite ordering: However this appears to be far from straightforward – the rank deficiency of $\hat{\Sigma}$ is the main technical issue to tackle.

References

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

U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 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, 1999.

R.I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 616–623. IEEE, 1999.

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

E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In F. Provost and R. Srikant, editor, *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, pages 245–250, 2001.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

G. Brown. *Encyclopedia of Machine Learning*, chapter Ensemble Learning. Springer, 2009.

G.C. Cawley. MATLAB support vector machine toolbox (v0.55β) University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ, 2000. URL http://theoval.cmp.uea.ac.uk/svm/toolbox/.

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

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457): 77–87, 2002.

R.J. Durrant and A. Kabán. Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data. In *Proceedings16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010.

R. Folgieri. *Ensembles based on Random Projection for gene expression data analysis*. PhD thesis, 2008. URL http://hdl.handle.net/2434/45878.

D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 522–529. ACM, 2003.

G. Fumera, F. Roli, and A. Serrau. A theoretical analysis of bagging as a linear combination of classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(7):1293–1299, 2008.

N. Goel, G. Bebis, and A. Nefian. Face recognition experiments with random projection. In *Proceedings of SPIE*, volume 5779, page 426, 2005.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531, 1999.

Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning; data mining, inference, and prediction*. Springer, 2001.

T.K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

T.L. Marzetta, G.H. Tucci, and S.H. Simon. A Random Matrix–Theoretic Approach to Handling Singular Covariance Estimates. *IEEE Trans. Information Theory*, 57(9):6256–71, September 2011.

S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and KR Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48. IEEE, 2002. ISBN 078035673X.

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

S. Raudys and R.P.W. Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19(5):385–392, 1998.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

A. Schclar and L. Rokach. Random projection ensemble classifiers. In Joaquim Filipe, Jos Cordeiro, Wil Aalst, John Mylopoulos, Michael Rosemann, Michael J. Shaw, and Clemens Szyperski, editors, *Enterprise Information Systems*, volume 24 of *Lecture Notes in Business Information Processing*, pages 309–316. Springer, 2009.

D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.S. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

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