ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 4, *pp.* 959–1035

Random-projection ensemble classification

Timothy I. Cannings and Richard J. Samworth

University of Cambridge, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 15th, 2017*, Professor C. Leng *in the Chair*]

Summary. We introduce a very general method for high dimensional classification, based on careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower dimensional space. In one special case that we study in detail, the random projections are divided into disjoint groups, and within each group we select the projection yielding the smallest estimate of the test error. Our random-projection ensemble classifier then aggregates the results of applying the base classifier on the selected projections, with a data-driven voting threshold to determine the final assignment. Our theoretical results elucidate the effect on performance of increasing the number of projections. Moreover, under a boundary condition that is implied by the sufficient dimension reduction assumption, we show that the test excess risk of the random-projection ensemble classifier can be controlled by terms that do not depend on the original data dimension and a term that becomes negligible as the number of projections increases. The classifier is also compared empirically with several other popular high dimensional classifiers via an extensive simulation study, which reveals its excellent finite sample performance.

Keywords: Aggregation; Classification; High dimensional classification; Random projection

1. Introduction

Supervised classification concerns the task of assigning an object (or a number of objects) to one of two or more groups, on the basis of a sample of labelled training data. The problem was first studied in generality in the famous work of Fisher (1936), where he introduced some of the ideas of linear discriminant analysis (LDA) and applied them to his iris data set. Nowadays, classification problems arise in a plethora of applications, including spam filtering, fraud detection, medical diagnoses, market research, natural language processing and many others.

In fact, LDA is still widely used today and underpins many other modern classifiers; see, for example, Friedman (1989) and Tibshirani *et al.* (2002). Alternative techniques include support vector machines (SVMs) (Cortes and Vapnik, 1995), tree classifiers and random forests (RFs) (Breiman *et al.*, 1984; Breiman, 2001), kernel methods (Hall and Kang, 2005) and nearest neighbour classifiers (Fix and Hodges, 1951). More substantial overviews and detailed discussion of these techniques, and others, can be found in Devroye *et al.* (1996) and Hastie *et al.* (2009).

An increasing number of modern classification problems are *high dimensional*, in the sense that the dimension p of the feature vectors may be comparable with or even greater than the number of training data points, n. In such settings, classical methods such as those mentioned in the previous paragraph tend to perform poorly (Bickel and Levina, 2004) and may even be

Address for correspondence: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: r.samworth@statslab.cam.ac.uk

1369–7412/17/79959

intractable; for example, this is so for LDA, where the problems are caused by the fact that the sample covariance matrix is not invertible when $p \geqslant n$.

Many methods proposed to overcome such problems assume that the optimal decision boundary between the classes is linear, e.g. Friedman (1989) and Hastie *et al.* (1995). Another common approach assumes that only a small subset of features are relevant for classification. Examples of works that impose such a sparsity condition include Fan and Fan (2008), where it was also assumed that the features are independent, as well as Tibshirani *et al.* (2003), where soft thresholding was used to obtain a sparse boundary. More recently, Witten and Tibshirani (2011) and Fan *et al.* (2012) both solved an optimization problem similar to Fisher's linear discriminant, with the addition of an l_1 penalty term to encourage sparsity.

In this paper we attempt to avoid the curse of dimensionality by projecting the feature vectors at random into a lower dimensional space. The use of random projections in high dimensional statistical problems is motivated by the celebrated Johnson–Lindenstrauss lemma (e.g. Dasgupta and Gupta (2002)). This lemma states that, given $x_1, \ldots, x_n \in \mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > 8\log(n)/\epsilon^2$, there is a linear map $f: \mathbb{R}^p \to \mathbb{R}^d$ such that

$$(1-\epsilon)\|x_i - x_j\|^2 \leqslant \|f(x_i) - f(x_j)\|^2 \leqslant (1+\epsilon)\|x_i - x_j\|^2,$$

for all $i, j = 1, \ldots, n$. In fact, the function f that nearly preserves the pairwise distances can be found in randomized polynomial time by using random projections distributed according to Haar measure, as described in Section 3 below. It is interesting to note that the lower bound on d in the Johnson–Lindenstrauss lemma does not depend on p; this lower bound is optimal up to constant factors (Larsen and Nelson, 2016). As a result, random projections have been used successfully as a computational time saver: when p is large compared with $\log(n)$, we may project the data at random into a lower dimensional space and run the statistical procedure on the projected data, potentially making great computational savings, while achieving comparable or even improved statistical performance. As one example of the above strategy, Durrant and Kabán (2013) obtained Vapnik–Chervonenkis-type bounds on the generalization error of a linear classifier trained on a single random projection of the data. See also Dasgupta (1999), Ailon and Chazelle (2006) and McWilliams *et al.* (2014) for other instances.

Other works have sought to reap the benefits of aggregating over many random projections. For instance, Marzetta *et al.* (2011) considered estimating a $p \times p$ population inverse covariance (precision) matrix by using $B^{-1}\Sigma_{b=1}^{B} \mathbf{A}_b^{\mathrm{T}}(\mathbf{A}_b\hat{\Sigma}\mathbf{A}_b^{\mathrm{T}})^{-1}\mathbf{A}_b$, where $\hat{\Sigma}$ denotes the sample covariance matrix and $\mathbf{A}_1, \ldots, \mathbf{A}_B$ are random projections from $\mathbb{R}^p$ to $\mathbb{R}^d$. Lopes *et al.* (2011) used this estimate when testing for a difference between two Gaussian population means in high dimensions, whereas Durrant and Kabán (2015) applied the same technique in Fisher's linear discriminant for a high dimensional classification problem.

Our proposed methodology for high dimensional classification has some similarities to the techniques described above, in the sense that we consider many random projections of the data, but is also closely related to *bagging* (Breiman, 1996), since the ultimate assignment of each test point is made by aggregation and a vote. Bagging has proved to be an effective tool for improving unstable classifiers. Indeed, a bagged version of the (generally inconsistent) 1-nearest-neighbour classifier is universally consistent as long as the resample size is carefully chosen: see Hall and Samworth (2005); for a general theoretical analysis of majority voting approaches, see also Lopes (2016). Bagging has also been shown to be particularly effective in high dimensional problems such as variable selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). Another related approach to ours is Blaser and Fryzlewicz (2015), who considered ensembles of random rotations, as opposed to projections.

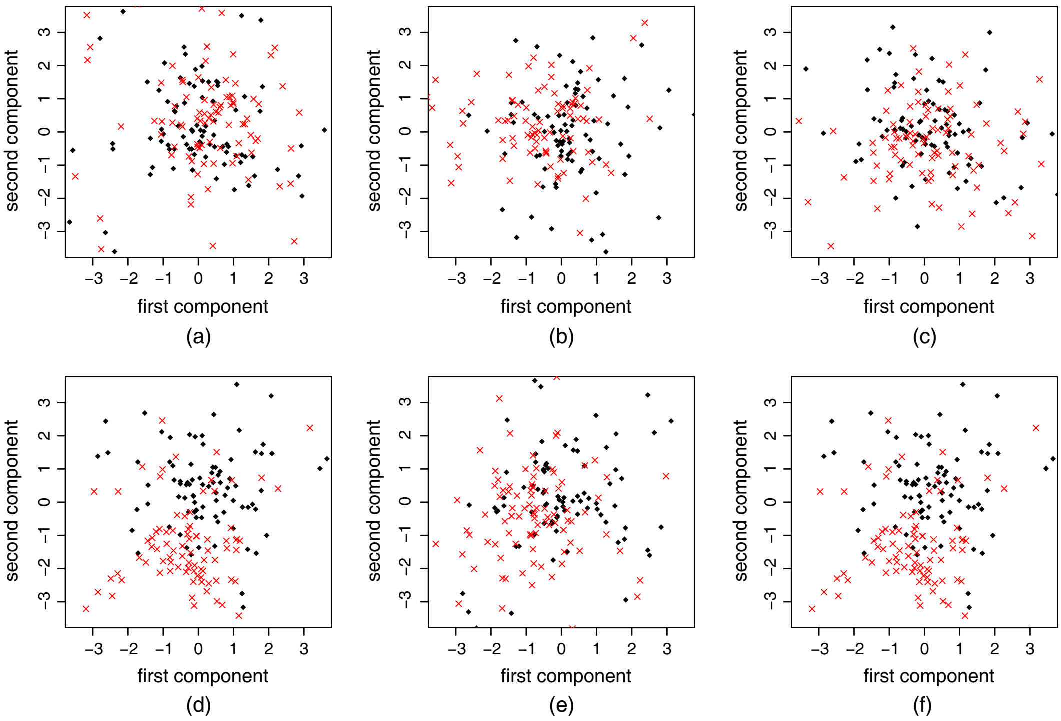


Fig. 1. Different two-dimensional projections of 200 observations in $p=50$ dimensions: (a)–(c) three projections drawn from Haar measure; (d)–(f) the projected data after applying the projections with smallest estimate of test error out of 100 Haar projections with (d) LDA, (e) quadratic discriminant analysis and (f) k-nearest neighbours

One of the basic but fundamental observations that underpins our proposal is the fact that aggregating the classifications of all random projections is not always sensible, since many of these projections will typically destroy the class structure in the data; see Figs 1(a)–1(c). For this reason, we advocate partitioning the projections into disjoint groups, and within each group we retain only the projection yielding the smallest estimate of the test error. The attraction of this strategy is illustrated in Figs 1(d)–1(f), where we see a much clearer partition of the classes. Another key feature of our proposal is the realization that a simple majority vote of the classifications based on the retained projections can be highly suboptimal; instead, we argue that the voting threshold should be chosen in a data-driven fashion in an attempt to minimize the test error of the infinite simulation version of our random-projection ensemble classifier. In fact, this estimate of the optimal threshold turns out to be remarkably effective in practice; see Section 5.2 for further details. We emphasize that our methodology can be used in conjunction with any base classifier, though we particularly have in mind classifiers designed for use in low dimensional settings. The random-projection ensemble classifier can therefore be regarded as a general technique for either extending the applicability of an existing classifier to high dimensions, or improving its performance. The methodology is implemented in an R package `RPEnsemble` (Cannings and Samworth, 2016).

Our theoretical results are divided into three parts. In the first, we consider a generic base classifier and a generic method for generating the random projections into $\mathbb{R}^d$ and quantify the difference between the test error of the random-projection ensemble classifier and its infinite simulation counterpart as the number of projections increases. We then consider selecting random projections from non-overlapping groups by initially drawing them according to Haar measure, and then within each group retaining the projection that minimizes an estimate of the test error. Under a condition that is implied by the widely used sufficient dimension reduction assumption (Li, 1991; Cook, 1998; Lee *et al.*, 2013), we can then control the difference between the test error of the random-projection classifier and the Bayes risk as a function of terms that depend on the performance of the base classifier based on projected data and our method for estimating the test error, as well as a term that becomes negligible as the number of projections increases. The final part of our theory gives risk bounds for the first two of these terms for specific choices of base classifier, namely Fisher's linear discriminant and the k-nearest-neighbour classifier. The key point here is that these bounds depend on d only, the sample size n and the number of projections, and not on the original data dimension p.

The remainder of the paper is organized as follows. Our methodology and general theory are developed in Sections 2 and 3. Specific choices of base classifier as well as a general sample splitting strategy are discussed in Section 4, whereas Section 5 is devoted to a consideration of the practical issues of computational complexity, choice of voting threshold, projected dimension and the number of projections used. In Section 6 we present results from an extensive empirical analysis on both simulated and real data where we compare the performance of the random-projection ensemble classifier with several popular techniques for high dimensional classification. The outcomes are very encouraging, and suggest that the random-projection ensemble classifier has excellent finite sample performance in a variety of high dimensional classification settings. We conclude with a discussion of various extensions and open problems. Proofs are given in Appendix A and the on-line supplementary material.

The program code that was used to perform the simulations can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

Finally in this section, we introduce the following general notation that is used throughout the paper. For a sufficiently smooth real-valued function g defined on a neighbourhood of $t \in \mathbb{R}$, let $\dot{g}(t)$ and $\ddot{g}(t)$ denote its first and second derivatives at t, and let $\lfloor t \rfloor$ and $[[t]] := t - \lfloor t \rfloor$ denote the integer and fractional part of t respectively.

2. A generic random-projection ensemble classifier

We start by describing our setting and defining the relevant notation. Suppose that the pair (X, Y) takes values in $\mathbb{R}^p \times \{0, 1\}$, with joint distribution P, characterized by $\pi_1 := \mathbb{P}(Y = 1)$, and P_r, the conditional distribution of $X|Y = r$, for $r = 0, 1$. For convenience, we let $\pi_0 := \mathbb{P}(Y = 0) = 1 - \pi_1$. In the alternative characterization of P, we let P_X denote the marginal distribution of X and write $\eta(x) := \mathbb{P}(Y = 1|X = x)$ for the regression function. Recall that a *classifier* on $\mathbb{R}^p$ is a Borel measurable function $C: \mathbb{R}^p \to \{0, 1\}$, with the interpretation that we assign a point $x \in \mathbb{R}^p$ to class $C(x)$. We let $\mathcal{C}_p$ denote the set of all such classifiers.

The test error of a classifier C is

$$R(C) := \int_{\mathbb{R}^p \times \{0,1\}} \mathbb{1}_{\{C(x) \neq y\}} \, \mathrm{d}P(x, y)$$

and is minimized by the *Bayes* classifier

$$C^{\mathrm{Bayes}}(x) := \begin{cases} 1 & \text{if } \eta(x) \geqslant \frac{1}{2}, \\ 0 & \text{otherwise} \end{cases}$$

(e.g. Devroye *et al.* (1996), page 10). (We define $R(C)$ through an integral rather than $R(C) :=$

$\mathbb{P}\{C(X) \neq Y\}$ to make it clear that, when C is random (depending on training data or random projections), it should be conditioned on when computing $R(C)$.) The Bayes risk is $R(C^{\text{Bayes}}) = \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}]$.

Of course, we cannot use the Bayes classifier in practice, since η is unknown. Nevertheless, we often have access to a sample of training data that we can use to construct an approximation to the Bayes classifier. Throughout this section and Section 3, it is convenient to consider the training sample $\mathcal{T}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to be fixed points in $\mathbb{R}^p \times \{0, 1\}$. Our methodology will be applied to a base classifier $C_n = C_{n,\mathcal{T}_{n,d}}$, which we assume can be constructed from an arbitrary training sample $\mathcal{T}_{n,d}$ of size n in $\mathbb{R}^d \times \{0, 1\}$; thus C_n is a measurable function from $(\mathbb{R}^d \times \{0, 1\})^n$ to $\mathcal{C}_d$.

Now assume that $d \leqslant p$. We say that a matrix $A \in \mathbb{R}^{d\times p}$ is a *projection* if $AA^{\mathrm{T}} = I_{d\times d}$, the d-dimensional identity matrix. Let $\mathcal{A} = \mathcal{A}_{d\times p} := \{A \in \mathbb{R}^{d\times p} : AA^{\mathrm{T}} = I_{d\times d}\}$ be the set of all such matrices. Given a projection $A \in \mathcal{A}$, define projected data $z_i^A := Ax_i$ and $y_i^A := y_i$ for $i = 1, \ldots, n$, and let $\mathcal{T}_n^A := \{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$. The projected data base classifier corresponding to C_n is $C_n^A : (\mathbb{R}^d \times \{0, 1\})^n \to \mathcal{C}_p$, given by

$$C_n^A(x) = C_{n,\mathcal{T}_n^A}^A(x) := C_{n,\mathcal{T}_n^A}(Ax).$$

Note that although C_n^A is a classifier on $\mathbb{R}^p$, the value of $C_n^A(x)$ only depends on x through its d-dimensional projection Ax.

We now define a generic ensemble classifier based on random projections. For $B_1 \in \mathbb{N}$, let $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$ denote independent and identically distributed projections in $\mathcal{A}_{d\times p}$, independent of (X, Y). The distribution on $\mathcal{A}$ is left unspecified at this stage, and in fact our proposed method ultimately involves choosing this distribution depending on $\mathcal{T}_n$.

Now set

$$\nu_n(x) = \nu_n^{(B_1)}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_n^{\mathbf{A}_{b_1}}(x)=1\}}. \tag{1}$$

For $\alpha \in (0, 1)$, the *random-projection ensemble* classifier is defined to be

$$C_n^{\mathrm{RP}}(x) := \begin{cases} 1 & \text{if } \nu_n(x) \geqslant \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We emphasize again here the additional flexibility that is afforded by not prespecifying the voting threshold α to be $\frac{1}{2}$. Our analysis of the random-projection ensemble classifier will require some further definitions. Let

$$\mu_n(x) := \mathbf{E}\{\nu_n(x)\} = \mathbf{P}\{C_n^{\mathbf{A}_1}(x) = 1\}.$$

(To distinguish between different sources of randomness, we shall write $\mathbf{P}$ and $\mathbf{E}$ for the probability and expectation respectively, taken over the randomness from the projections $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$. If the training data are random, then we condition on $\mathcal{T}_n$ when computing $\mathbf{P}$ and $\mathbf{E}$.) For $r = 0, 1$, define distribution functions $G_{n,r} : [0, 1] \to [0, 1]$ by $G_{n,r}(t) := P_r[\{x \in \mathbb{R}^p : \mu_n(x) \leqslant t\}]$. Since $G_{n,r}$ is non-decreasing it is differentiable almost everywhere; in fact, however, the following assumption will be convenient.

Assumption 1. $G_{n,0}$ and $G_{n,1}$ are twice differentiable at α.

The first derivatives of $G_{n,0}$ and $G_{n,1}$, when they exist, are denoted as $g_{n,0}$ and $g_{n,1}$ respectively; under assumption 1, these derivatives are well defined in a neighbourhood of α. Our first main result below gives an asymptotic expansion for the expected test error $\mathbf{E}\{R(C_n^{\mathrm{RP}})\}$ of our generic random-projection ensemble classifier as the number of projections increases. In particular, we show that this expected test error can be well approximated by the test error of the infinite simulation random-projection classifier

$$C_n^{\mathrm{RP}^*}(x) := \begin{cases} 1 & \text{if } \mu_n(x) \geqslant \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

Provided that $G_{n,0}$ and $G_{n,1}$ are continuous at α, we have

$$R(C_n^{\mathrm{RP}^*}) = \pi_1 G_{n,1}(\alpha) + \pi_0\{1 - G_{n,0}(\alpha)\}. \tag{3}$$

Theorem 1. Assume assumption 1. Then

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C_n^{\mathrm{RP}^*}) = \frac{\gamma_n(\alpha)}{B_1} + o\left(\frac{1}{B_1}\right)$$

as $B_1 \to \infty$, where

$$\gamma_n(\alpha) := (1 - \alpha - [[B_1\alpha]])\{\pi_1 g_{n,1}(\alpha) - \pi_0 g_{n,0}(\alpha)\} + \frac{\alpha(1-\alpha)}{2}\{\pi_1 \dot{g}_{n,1}(\alpha) - \pi_0 \dot{g}_{n,0}(\alpha)\}.$$

The proof of theorem 1 in Appendix A is lengthy and involves a one-term Edgeworth approximation to the distribution function of a standardized binomial random variable. One of the technical challenges is to show that the error in this approximation holds uniformly in the binomial proportion. Related techniques can also be used to show that $\mathbf{var}\{R(C_n^{\mathrm{RP}})\} = O(B_1^{-1})$ under assumption 1; see proposition 1 in the on-line supplementary material. Very recently, Lopes (2016) has obtained similar results to this and to theorem 1 in the context of majority vote classification, with stronger assumptions on the relevant distributions and on the form of the voting scheme. In Fig. 2, we plot the average error (±2 standard deviations) of the random-projection ensemble classifier in one numerical example, as we vary $B_1 \in \{2, \ldots, 500\}$; this reveals that the Monte Carlo error stabilizes rapidly, in agreement with what Lopes (2016) observed for a random-forest classifier.

Our next result controls the test excess risk, i.e. the difference between the expected test error and the Bayes risk, of the random-projection classifier in terms of the expected test excess risk of the classifier based on a single random projection. An attractive feature of this result is its generality: no assumptions are placed on the configuration of the training data $\mathcal{T}_n$, the distribution P of the test point (X, Y) or on the distribution of the individual projections.

Theorem 2. For each $B_1 \in \mathbb{N} \cup \{\infty\}$, we have

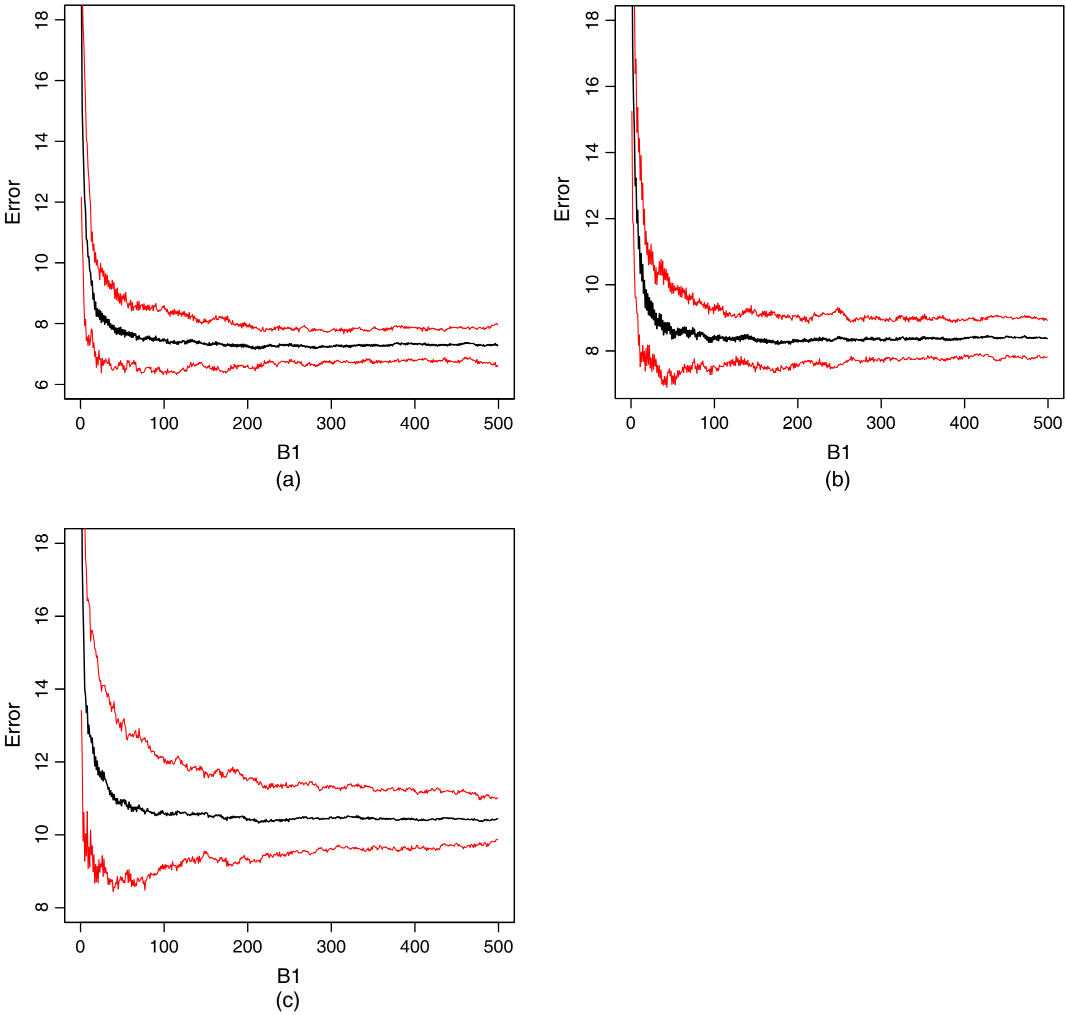
$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}}) \leqslant \frac{1}{\min(\alpha, 1-\alpha)}[\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\mathrm{Bayes}})]. \tag{4}$$

When $B_1 = \infty$, we interpret $R(C_n^{\mathrm{RP}})$ in theorem 2 as $R(C_n^{\mathrm{RP}^*})$. In fact, when $B_1 = \infty$ and $G_{n,0}$ and $G_{n,1}$ are continuous, the bound in theorem 2 can be improved if we are using an 'oracle' choice of the voting threshold α, namely

$$\alpha^* \in \underset{\alpha' \in [0,1]}{\arg\min}\, R(C_{n,\alpha'}^{\mathrm{RP}^*}) = \underset{\alpha' \in [0,1]}{\arg\min}[\pi_1 G_{n,1}(\alpha') + \pi_0\{1 - G_{n,0}(\alpha')\}], \tag{5}$$

where we write $C_{n,\alpha}^{\mathrm{RP}^*}$ to emphasize the dependence on the voting threshold α. In this case, by definition of α^* and then applying theorem 2,

$$R(C_{n,\alpha^*}^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) \leqslant R(C_{n,1/2}^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) \leqslant 2[\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\mathrm{Bayes}})], \tag{6}$$



(a)

(b)

(c)

Fig. 2. Average error (——) ±2 standard deviations (——) over 20 sets of $B_1 B_2$ projections for $B_1 \in \{2, \ldots, 500\}$: we use (a) the LDA, (b) quadratic discriminant analysis and (c) k-nearest-neighbour base classifiers (the plots show the test error for one training data set from model 2; the other parameters are $n = 50$, $p = 100$, $d = 5$ and $B_2 = 50$)

which improves the bound in expression (4) since $2 \leqslant 1/\min\{\alpha^*, (1-\alpha^*)\}$. It is also worth mentioning that if assumption 1 holds at $\alpha^* \in (0,1)$, and $G_{n,0}$ and $G_{n,1}$ are continuous, then $\pi_1 g_{n,1}(\alpha^*) = \pi_0 g_{n,0}(\alpha^*)$ and the constant in theorem 1 simplifies to

$$\gamma_n(\alpha^*) = \frac{\alpha^*(1-\alpha^*)}{2} \{\pi_1 \dot{g}_{n,1}(\alpha^*) - \pi_0 \dot{g}_{n,0}(\alpha^*)\} \geqslant 0.$$

3. Choosing good random projections

In this section, we study a special case of the generic random-projection ensemble classifier that was introduced in Section 2, where we propose a screening method for choosing the random projections. Let R_n^A be an estimator of $R(C_n^A)$, based on $\{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$, that takes values in the set $\{0, 1/n, \ldots, 1\}$. Examples of such estimators include the training error and leave-one-out estimator; we discuss these choices in greater detail in Section 4. For $B_1, B_2 \in \mathbb{N}$, let $\{\mathbf{A}_{b_1,b_2} : b_1 = 1, \ldots, B_1, b_2 = 1, \ldots, B_2\}$ denote independent projections, independent of (X, Y), distributed according to Haar measure on $\mathcal{A}$. One way to simulate from Haar measure on the set $\mathcal{A}$ is first to generate a matrix $\mathbf{Q} \in \mathbb{R}^{d\times p}$, where each entry is drawn independently from a standard normal distribution, and then to take $\mathbf{A}^{\mathrm{T}}$ to be the matrix of left singular vectors in the singular value decomposition of $\mathbf{Q}^{\mathrm{T}}$ (see, for example, Chikuse (2003), theorem 1.5.4). For $b_1 = 1, \ldots, B_1$, let

$$b_2^*(b_1) := \underset{b_2 \in \{1,\ldots,B_2\}}{\operatorname{sarg\,min}} R_n^{\mathbf{A}_{b_1,b_2}}, \tag{7}$$

where sargmin denotes the smallest index where the minimum is attained in the case of a tie. We now set $\mathbf{A}_{b_1} := \mathbf{A}_{b_1, b_2^*(b_1)}$, and consider the random-projection ensemble classifier from Section 2 constructed by using the independent projections $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$.

Let

$$R_n^* := \min_{A \in \mathcal{A}} R_n^A$$

denote the optimal test error estimate over all projections. The minimum is attained here, since R_n^A takes only finitely many values. We make the following assumption.

Assumption 2. There exists $\beta \in (0, 1]$ such that

$$\mathbf{P}(R_n^{\mathbf{A}_{1,1}} \leqslant R_n^* + |\epsilon_n|) \geqslant \beta,$$

where $\epsilon_n = \epsilon_n^{(B_2)} := \mathbf{E}\{R(C_n^{\mathbf{A}_1}) - R_n^{\mathbf{A}_1}\}$.

The quantity ϵ_n, which depends on B_2 because $\mathbf{A}_1$ is selected from B_2 independent random projections, can be interpreted as a measure of overfitting. Assumption 2 asks that there is a positive probability that $R_n^{\mathbf{A}_{1,1}}$ is within $|\epsilon_n|$ of its minimum value R_n^*. The intuition here is that spending more computational time choosing a projection by increasing B_2 is potentially futile: one may find a projection with a lower error estimate, but the chosen projection will not necessarily result in a classifier with a lower test error. Under this condition, the following result controls the test excess risk of our random-projection ensemble classifier in terms of the test excess risk of a classifier based on d-dimensional data, as well as a term that reflects our ability to estimate the test error of classifiers on the basis of projected data and a term that depends on the number of projections.

Theorem 3. Assume assumption 2. Then, for each $B_1, B_2 \in \mathbb{N}$, and every $A \in \mathcal{A}$,

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}}) \leqslant \frac{R(C_n^A) - R(C^{\mathrm{Bayes}})}{\min(\alpha, 1-\alpha)} + \frac{2|\epsilon_n| - \epsilon_n^A}{\min(\alpha, 1-\alpha)} + \frac{(1-\beta)^{B_2}}{\min(\alpha, 1-\alpha)}, \tag{8}$$

where $\epsilon_n^A := R(C_n^A) - R_n^A$.

Regarding the bound in theorem 3 as a sum of three terms, we see that the final term can be seen as the price that we must pay for the fact that we do not have access to an infinite sample of random projections. This term can be made negligible by choosing B_2 to be sufficiently large, though the value of B_2 that is required to ensure that it is below a prescribed level may depend on the training data. It should also be noted that ϵ_n in the second term may increase with B_2, which reflects the fact mentioned previously that this quantity is a measure of overfitting. The behaviour of the first two terms depends on the choice of base classifier, and our aim is to show that, under certain conditions, these terms can be bounded (in expectation over the training data) by expressions that do not depend on p.

For this, define the regression function on $\mathbb{R}^d$ induced by the projection $A \in \mathcal{A}$ to be $\eta^A(z) := \mathbb{P}(Y=1|AX=z)$. The corresponding induced Bayes classifier, which is the optimal classifier knowing only the distribution of (AX, Y), is given by

$$C^{A-\text{Bayes}}(z) := \begin{cases} 1 & \text{if } \eta^A(z) \geqslant \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

To give a condition under which there is a projection $A \in \mathcal{A}$ for which $R(C_n^A)$ is close to the Bayes risk, we shall invoke an additional assumption on the form of the Bayes classifier.

Assumption 3. There is a projection $A^* \in \mathcal{A}$ such that

$$P_X(\{x \in \mathbb{R}^p : \eta(x) \geqslant \tfrac{1}{2}\} \triangle \{x \in \mathbb{R}^p : \eta^{A^*}(A^*x) \geqslant \tfrac{1}{2}\}) = 0,$$

where $B \triangle C := (B \cap C^c) \cup (B^c \cap C)$ denotes the symmetric difference of two sets B and C.

Assumption 3 requires that the set of points $x \in \mathbb{R}^p$ that are assigned by the Bayes classifier to class 1 can be expressed as a function of a d-dimensional projection of x. If the Bayes decision boundary is a hyperplane, then assumption 3 holds with $d=1$. Moreover, proposition 1 below shows that, in fact, assumption 3 holds under the sufficient dimension reduction condition, which states that Y is conditionally independent of X given A^*X; see Cook (1998) for many statistical settings where such an assumption is natural.

Proposition 1. If Y is conditionally independent of X given A^*X, then assumption 3 holds.

The following result confirms that under assumption 3, and for a sensible choice of base classifier, we can hope for $R(C_n^{A^*})$ to be close to the Bayes risk.

Proposition 2. Assume assumption 3. Then $R(C^{A^*-\text{Bayes}}) = R(C^{\text{Bayes}})$.

We are therefore now ready to study the first two terms in the bound in theorem 3 in more detail for specific choices of base classifier.

4. Possible choices of the base classifier

In this section, we change our previous perspective and regard the training data as independent random pairs with distribution P, so our earlier statements are interpreted conditionally on $\mathcal{T}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. For $A \in \mathcal{A}$, we write our projected data as $\mathcal{T}_n^A := \{(Z_1^A, Y_1^A), \ldots, (Z_n^A, Y_n^A)\}$, where $Z_i^A := AX_i$ and $Y_i^A := Y_i$. We also write $\mathbb{P}$ and $\mathbb{E}$ to refer to probabilities and expectations over all random quantities. We consider particular choices of base classifier and study the first two terms in the bound in theorem 3.

4.1. Linear discriminant analysis

LDA, which was introduced by Fisher (1936), is arguably the simplest classification technique. Recall that, in the special case where $X|Y=r \sim N_p(\mu_r, \Sigma)$, we have

$$\text{sgn}\left\{\eta(x) - \frac{1}{2}\right\} = \text{sgn}\left\{\log\left(\frac{\pi_1}{\pi_0}\right) + \left(x - \frac{\mu_1+\mu_0}{2}\right)^{\mathrm{T}} \Sigma^{-1}(\mu_1 - \mu_0)\right\},$$

so assumption 3 holds with $d=1$ and $A^* = (\mu_1 - \mu_0)^{\mathrm{T}}\Sigma^{-1}/\|\Sigma^{-1}(\mu_1 - \mu_0)\|$, which is a $1 \times p$ matrix. In LDA, π_r, μ_r and Σ are estimated by their sample versions, using a pooled estimate of Σ. Although LDA cannot be applied directly when $p \geqslant n$ since the sample covariance matrix is singular, we can still use it as the base classifier for a random-projection ensemble, provided that $d < n$. Indeed, noting that, for any $A \in \mathcal{A}$, we have $AX|Y = r \sim N_d(\mu_r^A, \Sigma^A)$, where $\mu_r^A := A\mu_r$ and $\Sigma^A := A\Sigma A^{\mathrm{T}}$, we can define

$$C_n^A(x) = C_n^{A-\mathrm{LDA}}(x) := \begin{cases} 1 & \text{if } \log(\hat{\pi}_1/\hat{\pi}_0) + (Ax - (\hat{\mu}_1^A + \hat{\mu}_0^A)/2)^{\mathrm{T}} \hat{\Omega}^A (\hat{\mu}_1^A - \hat{\mu}_0^A) \geqslant 0; \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Here, $\hat{\pi}_r := n_r/n$, where $n_r := \Sigma_{i=1}^n \mathbb{1}_{\{Y_i = r\}}$, $\hat{\mu}_r^A := n_r^{-1} \Sigma_{i=1}^n AX_i \mathbb{1}_{\{Y_i = r\}}$,

$$\hat{\Sigma}^A := \frac{1}{n-2} \sum_{i=1}^n \sum_{r=0}^1 (AX_i - \hat{\mu}_r^A)(AX_i - \hat{\mu}_r^A)^{\mathrm{T}} \mathbb{1}_{\{Y_i = r\}}$$

and $\hat{\Omega}^A := (\hat{\Sigma}^A)^{-1}$.

Write Φ for the standard normal distribution function. Under the normal model specified above, the test error of the LDA classifier can be written as

$$R(C_n^A) = \pi_0 \, \Phi\left[\frac{\log(\hat{\pi}_1/\hat{\pi}_0) + (\hat{\delta}^A)^{\mathrm{T}} \hat{\Omega}^A (\bar{\hat{\mu}}^A - \mu_0^A)}{\surd\{(\hat{\delta}^A)^{\mathrm{T}} \hat{\Omega}^A \Sigma^A \hat{\Omega}^A \hat{\delta}^A\}} \right] + \pi_1 \, \Phi\left[\frac{\log(\hat{\pi}_0/\hat{\pi}_1) - (\hat{\delta}^A)^{\mathrm{T}} \hat{\Omega}^A (\bar{\hat{\mu}}^A - \mu_1^A)}{\surd\{(\hat{\delta}^A)^{\mathrm{T}} \hat{\Omega}^A \Sigma^A \hat{\Omega}^A \hat{\delta}^A\}} \right],$$

where $\hat{\delta}^A := \hat{\mu}_0^A - \hat{\mu}_1^A$ and $\bar{\hat{\mu}}^A := (\hat{\mu}_0^A + \hat{\mu}_1^A)/2$.

Efron (1975) studied the excess risk of the LDA classifier in an asymptotic regime in which d is fixed as n diverges. Specializing his results for simplicity to the case where $\pi_0 = \pi_1$, he showed that using the LDA classifier (9) with $A = A^*$ yields

$$\mathbb{E}\{R(C_n^{A^*})\} - R(C^{\mathrm{Bayes}}) = \frac{d}{n} \phi\left(-\frac{\Delta}{2}\right) \left(\frac{\Delta}{4} + \frac{1}{\Delta}\right) \{1 + o(1)\} \tag{10}$$

as $n \to \infty$, where $\Delta := \|\Sigma^{-1/2}(\mu_0 - \mu_1)\| = \|(\Sigma^{A^*})^{-1/2}(\mu_0^{A^*} - \mu_1^{A^*})\|$.

It remains to control the errors ϵ_n and $\epsilon_n^{A^*}$ in theorem 3. For the LDA classifier, we consider the training error estimator

$$R_n^A := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{C_n^{A-\mathrm{LDA}}(X_i) \neq Y_i\}}. \tag{11}$$

Devroye and Wagner (1976) provided a Vapnik–Chervonenkis bound for R_n^A under no assumptions on the underlying data-generating mechanism: for every $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\sup_{A \in \mathcal{A}} \mathbb{P}\{|R(C_n^A) - R_n^A| > \epsilon\} \leqslant 8(n+1)^{d+1} \exp(-n\epsilon^2/32); \tag{12}$$

see also Devroye *et al.* (1996), theorem 23.1. We can then conclude that

$$\begin{aligned} \mathbb{E}|\epsilon_n^{A^*}| \leqslant \mathbb{E}|R(C_n^{A^*}) - R_n^{A^*}| &\leqslant \inf_{\epsilon_0 \in (0,1)} \epsilon_0 + 8(n+1)^{d+1} \int_{\epsilon_0}^1 \exp\left(-\frac{ns^2}{32}\right) \mathrm{d}s \\ &\leqslant 8 \sqrt{\left\{\frac{(d+1)\log(n+1) + 3\log(2) + 1}{2n}\right\}}. \end{aligned} \tag{13}$$

The more difficult term to deal with is

$$\mathbb{E}|\epsilon_n| = \mathbb{E}|\mathbf{E}\{R(C_n^{\mathbf{A}_1}) - R_n^{\mathbf{A}_1}\}| \leqslant \mathbb{E}|R(C_n^{\mathbf{A}_1}) - R_n^{\mathbf{A}_1}|.$$

In this case, the bound (12) cannot be applied directly, because $(X_1, Y_1), \ldots, (X_n, Y_n)$ are no longer independent conditional on $\mathbf{A}_1$; indeed $\mathbf{A}_1 = \mathbf{A}_{1, b_2^*(1)}$ is selected from $\mathbf{A}_{1,1}, \ldots, \mathbf{A}_{1,B_2}$ to minimize an estimate of test error, which depends on the training data. Nevertheless, since $\mathbf{A}_{1,1},\dots,\mathbf{A}_{1,B_2}$ are independent of $\mathcal{T}_n$, we still have that

$$\begin{aligned}\mathbb{P}\Big\{\max_{b_2=1,\dots,B_2} |R(C_n^{\mathbf{A}_{1,b_2}}) - R_n^{\mathbf{A}_{1,b_2}}| > \epsilon | \mathbf{A}_{1,1},\dots,\mathbf{A}_{1,B_2}\Big\} &\leqslant \sum_{b_2=1}^{B_2} \mathbb{P}\{|R(C_n^{\mathbf{A}_{1,b_2}}) - R_n^{\mathbf{A}_{1,b_2}}| > \epsilon | \mathbf{A}_{1,b_2}\} \\ &\leqslant 8(n+1)^{d+1} B_2 \exp(-n\epsilon^2/32).\end{aligned}$$

We can therefore conclude by almost the same argument as that leading to bound (13) that

$$\mathbb{E}|\epsilon_n| \leqslant \mathbb{E}\Big\{\max_{b_2=1,\dots,B_2} |R(C_n^{\mathbf{A}_{1,b_2}}) - R_n^{\mathbf{A}_{1,b_2}}|\Big\} \leqslant 8\sqrt{\left\{\frac{(d+1)\log(n+1) + 3\log(2) + \log(B_2) + 1}{2n}\right\}}. \tag{14}$$

Note that none of the bounds (10), (13) and (14) depend on the original data dimension p. Moreover, bound (14), together with theorem 3, reveals a trade-off in the choice of B_2 when using LDA as the base classifier. Choosing B_2 to be large gives us a good chance of finding a projection with a small estimate of test error, but we may incur a small overfitting penalty as reflected by bound (14).

Finally, we remark that an alternative method of fitting linear classifiers is via empirical risk minimization. In this context, Durrant and Kabán (2013), theorem 3.1, gave high probability bounds on the test error of a linear empirical risk minimization classifier based on a single random projection, where the bounds depend on what they refered to as the 'flipping probability', namely the probability that the class assignment of a point based on the projected data differs from the assignment in the ambient space. In principle, these bounds could be combined with our theorem 2, though the resulting expressions would depend on probabilistic bounds on flipping probabilities.

4.2. Quadratic discriminant analysis

Quadratic discriminant analysis (QDA) is designed to handle situations where the class conditional covariance matrices are unequal. Recall that when $X|Y=r \sim N_p(\mu_r, \Sigma_r)$, and $\pi_r = \mathbb{P}(Y = r)$, for $r = 0, 1$, the Bayes decision boundary is given by $\{x \in \mathbb{R}^p : \Delta(x; \pi_0, \mu_0, \mu_1, \Sigma_0, \Sigma_1) = 0\}$, where

$$\begin{aligned}\Delta(x; \pi_0, \mu_0, \mu_1, \Sigma_0, \Sigma_1) := \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\log\left\{\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right\} - \frac{1}{2}x^{\mathrm{T}}(\Sigma_1^{-1} - \Sigma_0^{-1})x \\ + x^{\mathrm{T}}(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) - \frac{1}{2}\mu_1^{\mathrm{T}}\Sigma_1^{-1}\mu_1 + \frac{1}{2}\mu_0^{\mathrm{T}}\Sigma_0^{-1}\mu_0.\end{aligned}$$

In QDA, π_r, μ_r and Σ_r are estimated by their sample versions. If $p \geqslant \min(n_0, n_1)$, where we recall that $n_r := \Sigma_{i=1}^n \mathbb{1}_{\{Y_i = r\}}$, then at least one of the sample covariance matrix estimates is singular, and QDA cannot be used directly. Nevertheless, we can still choose $d < \min(n_0, n_1)$ and use QDA as the base classifier in a random-projection ensemble. Specifically, we can set

$$C_n^A(x) = C_n^{A-\mathrm{QDA}}(x) := \begin{cases} 1 & \text{if } \Delta(x; \hat{\pi}_0, \hat{\mu}_0^A, \hat{\mu}_1^A, \hat{\Sigma}_0^A, \hat{\Sigma}_1^A) \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{\pi}_r$ and $\hat{\mu}_r^A$ were defined in Section 4.1, and where

$$\hat{\Sigma}_r^A := \frac{1}{n_r - 1} \sum_{\{i: Y_i^A = r\}} (AX_i - \hat{\mu}_r^A)(AX_i - \hat{\mu}_r^A)^{\mathrm{T}}$$

for $r = 0, 1$. Unfortunately, analogous theory to that presented in Section 4.1 does not appear to exist for the QDA classifier; unlike for LDA, the risk does not have a closed form (note that $\Sigma_1 - \Sigma_0$ is non-definite in general). Nevertheless, we found in our simulations presented in Section 6 that the QDA random-projection ensemble classifier can perform very well in practice. In this case, we estimate the test errors by using the leave-one-out method given by

$$R_n^A := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{C_{n,i}^A(X_i) \neq Y_i\}}, \tag{15}$$

where $C_{n,i}^A$ denotes the classifier C_n^A, trained without the ith pair, i.e. based on $\mathcal{T}_n^A \setminus \{Z_i^A, Y_i^A\}$. For a method like QDA that involves estimating more parameters than LDA, we found that the leave-one-out estimator was less susceptible to overfitting than the training error estimator.

4.3. The k-nearest-neighbour classifier

The k-nearest-neighbour classifier $k$nn, which was first proposed by Fix and Hodges (1951), is a non-parametric method that classifies the test point $x \in \mathbb{R}^p$ according to a majority vote over the classes of the k nearest training data points to x. The enormous popularity of $k$nn can be attributed partly to its simplicity and intuitive appeal; however, it also has the attractive property of being universally consistent: for every fixed distribution P, as long as $k \to \infty$ and $k/n \to 0$, the risk of $k$nn converges to the Bayes risk (Devroye *et al.* (1996), theorem 6.4).

Hall *et al.* (2008) studied the rate of convergence of the excess risk of the k-nearest-neighbour classifier under regularity conditions that require, *inter alia*, that p is fixed and that the class conditional densities have two continuous derivatives in a neighbourhood of the $(p-1)$-dimensional manifold on which they cross. In such settings, the optimal choice of k, in terms of minimizing the excess risk, is $O(n^{4/(p+4)})$, and the rate of convergence of the excess risk with this choice is $O(n^{-4/(p+4)})$. Thus, in common with other non-parametric methods, there is a 'curse-of-dimensionality' effect that means that the classifier typically performs poorly in high dimensions. Samworth (2012) found the optimal way of assigning decreasing weights to increasingly distant neighbours and quantified the asymptotic improvement in risk over the unweighted version, but the rate of convergence remains the same.

We can use $k$nn as the base classifier for a random-projection ensemble as follows: given $z \in \mathbb{R}^d$, let $(Z_{(1)}^A, Y_{(1)}^A), \ldots, (Z_{(n)}^A, Y_{(n)}^A)$ be a reordering of the training data such that $\|Z_{(1)}^A - z\| \leqslant \ldots \leqslant \|Z_{(n)}^A - z\|$, with ties split at random. Now define

$$C_n^A(x) = C_n^{A-k\text{nn}}(x) := \begin{cases} 1 & \text{if } S_n^A(Ax) \geqslant \frac{1}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

where $S_n^A(z) := k^{-1} \Sigma_{i=1}^k \mathbb{1}_{\{Y_{(i)}^A = 1\}}$. The theory that was described in the previous paragraph can be applied to show that, under regularity conditions, $\mathbb{E}\{R(C_n^{A^*})\} - R(C^{\text{Bayes}}) = O(n^{-4/(d+4)})$.

Once again, a natural estimate of the test error in this case is the leave-one-out estimator that is defined in expression (15), where we use the same value of k on the leave-one-out data sets as for the base classifier, and where distance ties are split in the same way as for the base classifier. For this estimator, Devroye and Wagner (1979), theorem 4, showed that, for every $n \in \mathbb{N}$,

$$\sup_{A \in \mathcal{A}} \mathbb{E}[\{R(C_n^A) - R_n^A\}^2] \leqslant \frac{1}{n} + \frac{24k^{1/2}}{n\sqrt{(2\pi)}};$$

see also Devroye *et al.* (1996), chapter 24. It follows that

$$\mathbb{E}|\epsilon_n^{A^*}| \leqslant \left\{\frac{1}{n} + \frac{24k^{1/2}}{n\surd(2\pi)}\right\}^{1/2} \leqslant \frac{1}{n^{1/2}} + \frac{2^{5/4}k^{1/4}\surd 3}{n^{1/2}\pi^{1/4}}.$$

In fact, Devroye and Wagner (1979), theorem 1, also provided a tail bound analogous to bound (12) for the leave-one-out estimator: for every $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\sup_{A\in\mathcal{A}} \mathbb{P}\{|R(C_n^A) - R_n^A| > \epsilon\} \leqslant 2\exp\left(-\frac{n\epsilon^2}{18}\right) + 6\exp\left\{-\frac{n\epsilon^3}{108k(3^d+1)}\right\} \leqslant 8\exp\left\{-\frac{n\epsilon^3}{108k(3^d+1)}\right\}.$$

Arguing very similarly to Section 4.1, we can deduce that under no conditions on the data-generating mechanism, and, choosing

$$\epsilon_0 := \left\{\frac{108k(3^d+1)}{n}\log(8B_2)\right\}^{1/3},$$

we have

$$\mathbb{E}|\epsilon_n| = \int_0^1 \mathbb{P}\left\{\max_{b_2=1,\ldots,B_2} |R(C_n^{\mathbf{A}_{1,b_2}}) - R_n^{\mathbf{A}_{1,b_2}}| > \epsilon\right\}\mathrm{d}\epsilon$$
$$\leqslant \epsilon_0 + 8B_2\int_{\epsilon_0}^{\infty} \exp\left\{-\frac{n\epsilon^3}{108k(3^d+1)}\right\}\mathrm{d}\epsilon \leqslant 3\{4(3^d+1)\}^{1/3}\left[\frac{k\{1+\log(B_2)+3\log(2)\}}{n}\right]^{1/3}.$$

We have therefore again bounded the expectations of the first two terms on the right-hand side of inequality (8) by quantities that do not depend on p.

4.4. A general strategy using sample splitting

In Sections 4.1, 4.2 and 4.3, we focused on specific choices of the base classifier to derive bounds on the expected values of the first two terms in the bound in theorem 3. The aim of this section, in contrast, is to provide similar guarantees that can be applied for any choice of base classifier in conjunction with a sample splitting strategy. The idea is to split the sample $\mathcal{T}_n$ into $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$, say, where $|\mathcal{T}_{n,1}| =: n^{(1)}$ and $|\mathcal{T}_{n,2}| =: n^{(2)}$. To estimate the test error of $C_{n^{(1)}}^A$, the projected data base classifier trained on $\mathcal{T}_{n,1}^A := \{(Z_i^A, Y_i^A) : (X_i, Y_i) \in \mathcal{T}_{n,1}\}$, we use

$$R_{n^{(1)},n^{(2)}}^A := \frac{1}{n^{(2)}} \sum_{(X_i,Y_i)\in\mathcal{T}_{n,2}} \mathbb{1}_{\{C_{n^{(1)}}^A(X_i)\neq Y_i\}};$$

in other words, we construct the classifier based on the projected data from $\mathcal{T}_{n,1}$ and count the proportion of points in $\mathcal{T}_{n,2}$ that are misclassified. Similarly to our previous approach, for the b_1th group of projections, we then select a projection $\mathbf{A}_{b_1}$ that minimizes this estimate of test error and construct the random-projection ensemble classifier $C_{n^{(1)},n^{(2)}}^{\mathrm{RP}}$ from

$$\nu_{n^{(1)}}(x) := \frac{1}{B_1}\sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_{n^{(1)}}^{\mathbf{A}_{b_1}}(x)=1\}}.$$

Writing $R_{n^{(1)},n^{(2)}}^* := \min_{A\in\mathcal{A}} R_{n^{(1)},n^{(2)}}^A$, we introduce the following assumption which is analogous to assumption 2.

Assumption 2′. There exists $\beta \in (0,1]$ such that

$$\mathbf{P}(R_{n^{(1)},n^{(2)}}^{\mathbf{A}_{1,1}} \leqslant R_{n^{(1)},n^{(2)}}^* + |\epsilon_{n^{(1)},n^{(2)}}|) \geqslant \beta,$$

where $\epsilon_{n^{(1)},n^{(2)}} := \mathbf{E}\{R(C_{n^{(1)}}^{\mathbf{A}_1}) - R_{n^{(1)},n^{(2)}}^{\mathbf{A}_1}\}$.

The following bound for the random-projection ensemble classifier with sample splitting is then immediate from theorem 3 and proposition 2.

Corollary 1. Assume assumptions 2′ and 3. Then, for each $B_1, B_2 \in \mathbb{N}$,

$$\mathbf{E}\{R(C^{\mathrm{RP}}_{n^{(1)},n^{(2)}})\} - R(C^{\mathrm{Bayes}}) \leqslant \frac{R(C^{A^*}_{n^{(1)}}) - R(C^{A^*\text{-Bayes}})}{\min(\alpha, 1-\alpha)} + \frac{2|\epsilon_{n^{(1)},n^{(2)}}| - \epsilon^{A^*}_{n^{(1)},n^{(2)}}}{\min(\alpha, 1-\alpha)} + \frac{(1-\beta)^{B_2}}{\min(\alpha, 1-\alpha)},$$

where $\epsilon^{A^*}_{n^{(1)},n^{(2)}} := R(C^{A^*}_{n^{(1)}}) - R^{A^*}_{n^{(1)},n^{(2)}}$.

The main advantage of this approach is that, since $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$ are independent, we can apply Hoeffding's inequality to deduce that

$$\sup_{A\in\mathcal{A}} \mathbb{P}\{|R(C^{A}_{n^{(1)}}) - R^{A}_{n^{(1)},n^{(2)}}| \geqslant \epsilon | \mathcal{T}_{n,1}\} \leqslant 2\exp(-2n^{(2)}\epsilon^2).$$

It then follows by very similar arguments to those given in Section 4.1 that

$$\begin{aligned} \mathbb{E}(|\epsilon^{A^*}_{n^{(1)},n^{(2)}}||\mathcal{T}_{n,1}) &= \mathbb{E}\{|R(C^{A^*}_{n^{(1)}}) - R^{A^*}_{n^{(1)},n^{(2)}}||\mathcal{T}_{n,1}\} \leqslant \left\{\frac{1+\log(2)}{2n^{(2)}}\right\}^{1/2}, \\ \mathbb{E}(|\epsilon_{n^{(1)},n^{(2)}}||\mathcal{T}_{n,1}) &= \mathbb{E}\{|R(C^{\mathbf{A}_1}_{n^{(1)}}) - R^{\mathbf{A}_1}_{n^{(1)},n^{(2)}}||\mathcal{T}_{n,1}\} \leqslant \left\{\frac{1+\log(2)+\log(B_2)}{2n^{(2)}}\right\}^{1/2}. \end{aligned} \tag{16}$$

These bounds hold for any choice of base classifier (and still without any assumptions on the data-generating mechanism); moreover, since the bounds on the terms in expression (16) merely rely on Hoeffding's inequality as opposed to Vapnik–Chervonenkis theory, they are typically sharper. The disadvantage is that the first term in the bound in corollary 1 will typically be larger than the corresponding term in theorem 3 because of the reduced effective sample size.

5. Practical considerations

5.1. Computational complexity

The random-projection ensemble classifier aggregates the results of applying a base classifier to many random projections of the data. Thus we need to compute the training error (or leave-one-out error) of the base classifier after applying each of the $B_1 B_2$ projections. The test point is then classified by using the B_1 projections that yield the minimum error estimate within each block of size B_2.

Generating a random projection from Haar measure involves computing the left singular vectors of a $p \times d$ matrix, which requires $O(pd^2)$ operations (Trefethen and Bau (1997), lecture 31). However, if computational cost is a concern, one may simply generate a matrix with pd independent $N(0, 1/p)$ entries. If p is large, such a matrix will be approximately orthonormal with high probability. In fact, when the base classifier is affine invariant (as for LDA and QDA), this will give the same results as using Haar projections, in which case one can forgo the orthonormalization step altogether when generating the random projections. Even in very high dimensional settings, multiplication by a random Gaussian matrix can be approximated in a computationally efficient manner (e.g. Le *et al.* (2013)). Once a projection has been generated, we need to map the training data to the lower dimensional space, which requires $O(npd)$ operations. We then classify the training data by using the base classifier. The cost of this step, which is denoted T_{train}, depends on the choice of the base classifier; for LDA and QDA it is $O(nd^2)$, whereas for $k$nn it is $O(n^2 d)$. Finally the test points are classified by using the chosen projections; this cost, which is denoted T_{test}, also depends on the choice of base classifier. For LDA, QDA and $k$nn it is

$O(n_{\mathrm{test}}d)$, $O\{n_{\mathrm{test}}d^2\}$ and $O\{(n+n_{\mathrm{test}})^2d\}$ respectively, where n_{test} is the number of test points. Overall we therefore have a cost of $O[B_1\{B_2(npd+T_{\mathrm{train}})+n_{\mathrm{test}}pd+T_{\mathrm{test}}\}]$ operations.

An appealing aspect of the proposal that is studied here is the scope to incorporate parallel computing. We can simultaneously compute the projected data base classifier for each of the B_1B_2 projections. In the on-line supplementary material we present the running times of the random-projection ensemble classifier and the other methods that are considered in the empirical comparison in Section 6.

5.2. Choice of α

We now discuss the choice of the voting threshold α. In equation (5), at the end of Section 2, we defined the oracle choice α^*, which minimizes the test error of the infinite simulation random-projection classifier. Of course, α^* cannot be used directly, because we do not know $G_{n,0}$ and $G_{n,1}$ (and we may not know π_0 and π_1 either). Nevertheless, for the LDA base classifier we can estimate $G_{n,r}$ by using

$$\hat{G}_{n,r}(t):=\frac{1}{n_r}\sum_{\{i:Y_i=r\}}\mathbb{1}_{\{\nu_n(X_i)<t\}}$$

for $r=0,1$. For the QDA and k-nearest-neighbour base classifiers, we use the leave-one-out based estimate

$$\tilde{\nu}_n(X_i):=B_1^{-1}\sum_{b_1=1}^{B_1}\mathbb{1}_{\{C_{n,i}^{A_{b_1}}(X_i)=1\}}$$

in place of $\nu_n(X_i)$. We also estimate π_r by $\hat{\pi}_r:=n^{-1}\Sigma_{i=1}^n\mathbb{1}_{\{Y_i=r\}}$, and then set the cut-off in definition (2) as

$$\hat{\alpha}\in\underset{\alpha'\in[0,1]}{\arg\min}[\hat{\pi}_1\hat{G}_{n,1}(\alpha')+\hat{\pi}_0\{1-\hat{G}_{n,0}(\alpha')\}]. \tag{17}$$

Since empirical distribution functions are piecewise constant, the objective function in expression (17) does not have a unique minimum, so we choose $\hat{\alpha}$ to be the average of the smallest and largest minimizers. An attractive feature of the method is that $\{\nu_n(X_i):i=1,\ldots,n\}$ (or $\{\tilde{\nu}_n(X_i):i=1,\ldots,n\}$ in the case of QDA and $k$nn) are already calculated in order to choose the projections, so calculating $\hat{G}_{n,0}$ and $\hat{G}_{n,1}$ carries negligible extra computational cost.

Fig. 3 illustrates $\hat{\pi}_1\hat{G}_{n,1}(\alpha')+\hat{\pi}_0\{1-\hat{G}_{n,0}(\alpha')\}$ as an estimator of $\pi_1G_{n,1}(\alpha')+\pi_0\{1-G_{n,0}(\alpha')\}$, for the QDA base classifier and for various values of n and π_1. Here, a very good approximation to the estimand was obtained by using an independent data set of size 5000. Unsurprisingly, the performance of the estimator improves as n increases, but the most notable feature of these plots is the fact that, for all classifiers and even for small sample sizes, $\hat{\alpha}$ is an excellent estimator of α^* and may be a substantial improvement on the naive choice $\hat{\alpha}=\frac{1}{2}$ (or the appropriate prior weighted choice), which may result in a classifier that assigns every point to a single class.

5.3. Choice of B_1 and B_2

To minimize the Monte Carlo error as described in theorem 1 and proposition 1, we should choose B_1 to be as large as possible. The constraint, of course, is that the computational cost of the random-projection classifier scales linearly with B_1. The choice of B_2 is more subtle; whereas the third term in the bound in theorem 3 decreases as B_2 increases, we saw in Section 4 that upper bounds on $\mathbb{E}|\epsilon_n|$ may increase with B_2. In principle, we could try to use the expressions

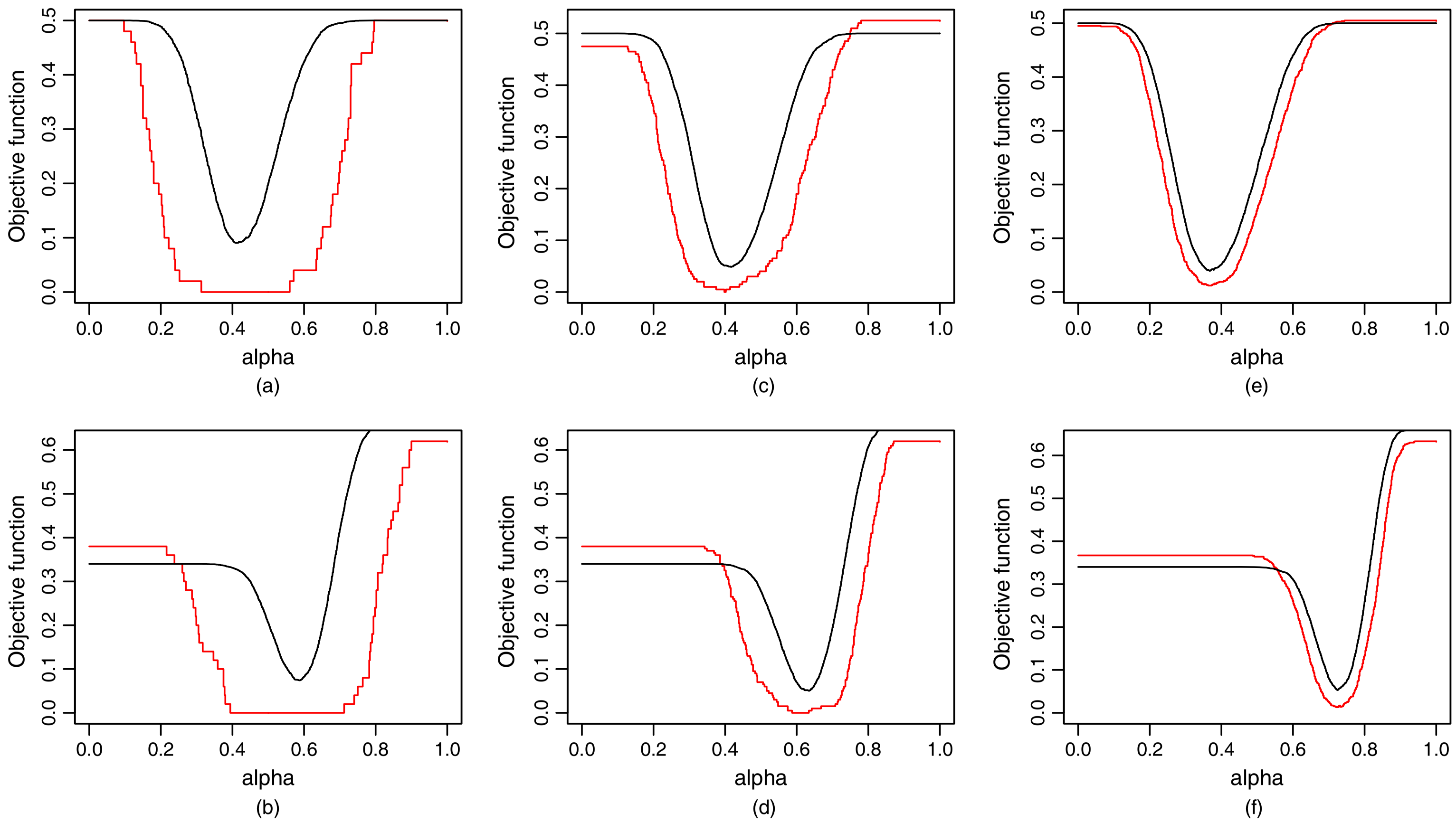


Fig. 3. $\pi_1 G_{n,1}(\alpha') + \pi_0\{1 - G_{n,0}(\alpha')\}$ in expression (5) (——) and $\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_0\{1 - \hat{G}_{n,0}(\alpha')\}$ (——) for the QDA base classifier after projecting for one training data set of size (a), (b) $n = 50$, (c), (d) $n = 200$ and (e), (f) $n = 1000$ from model 3 with (a), (c), (e) $\pi_1 = 0.5$ and (b), (d), (f) $\pi_1 = 0.66$: here, $p = 100$ and $d = 2$

that are given in theorem 3 and Section 4 to choose B_2 to minimize the overall upper bound on $\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}})$. In practice, however, we found that an involved approach such as this was unnecessary, and that the ensemble method was robust to the choice of B_1 and B_2; see Section 3 of the on-line supplementary material for numerical evidence and further discussion. On the basis of this numerical work, we recommend $B_1 = 500$ and $B_2 = 50$ as sensible default choices, and indeed these values were used in all of our experiments in Section 6 as well as section 4 in the supplementary material.

5.4. Choice of d

We want to choose d as small as possible to obtain the best possible performance bounds as described in Section 4. This also reduces the computational cost. However, the performance bounds rely on assumption 3, whose strength decreases as d increases, so we want to choose d sufficiently large that this condition holds (at least approximately).

In Section 6 we see that the random-projection ensemble method is quite robust to the choice of d. Nevertheless, in some circumstances it may be desirable to have an automatic choice, and cross-validation provides one possible approach when computational cost at *training time* is not too constrained. Thus, if we wish to choose d from a set $\mathcal{D} \subseteq \{1, \ldots, p\}$, then for each $d \in \mathcal{D}$ we train the random-projection ensemble classifier and set

$$\hat{d} := \underset{d \in \mathcal{D}}{\mathrm{sarg\,min}} [\hat{\pi}_1 \hat{G}_{n,1}(\hat{\alpha}) + \hat{\pi}_0\{1 - \hat{G}_{n,0}(\hat{\alpha})\}],$$

where $\hat{\alpha} = \hat{\alpha}_d$ is given in expression (17). Such a procedure does not add to the computational cost at *test time*. This strategy is most appropriate when $\max\{d : d \in \mathcal{D}\}$ is not too large (which is the setting that we have in mind); otherwise a penalized risk approach may be more suitable.

6. Empirical analysis

In this section, we assess the empirical performance of the random-projection ensemble classifier in simulated and real data experiments. We shall write RP-LDA$_d$, RP-QDA$_d$ and RP-$k$nn$_d$ to denote the random-projection classifier with LDA, QDA and $k$nn respectively; the subscript d refers to the dimension of the image space of the projections.

For comparison we present the corresponding results of applying, where possible, the three base classifiers (LDA, QDA, $k$nn) in the original p-dimensional space alongside 11 other classification methods chosen to represent the state of the art. These include RFs (Breiman, 2001); SVMs (Cortes and Vapnik, 1995), Gaussian process (GP) classifiers (Williams and Barber, 1998) and three methods designed for high dimensional classification problems, namely penalized LDA, PenLDA (Witten and Tibshirani, 2011), nearest shrunken centroids (NSCs) (Tibshirani *et al.*, 2003) and l_1-penalized logistic regression, PenLog (Goeman *et al.*, 2015).

A further comparison is with LDA and $k$nn applied after a single projection chosen on the basis of the sufficient dimension reduction assumption, SDR5. For this method, we project the data into five dimensions by using the proposal of Shin *et al.* (2014). This method requires $n > p$. Finally, we compare with two related ensemble methods: optimal tree ensembles (OTEs) (Khan *et al.*, 2015a) and ensemble of subset of k-nearest-neighbour classifiers, ES$k$nn (Gul *et al.*, 2016).

Many of these methods require tuning parameter selection, and the parameters were chosen as follows: for the $k$nn standard classifier, we chose k via leave-one-out cross-validation from the set $\{3, 5, 7, 9, 11\}$. RF was implemented by using the `randomForest` package (Liaw and Wiener, 2014); we used an ensemble of 1000 trees, with $\lfloor \sqrt{p} \rfloor$ (the default setting in the `randomForest` package) components randomly selected when training each tree. For the radial SVM, we used the reproducing basis kernel $K(u, v) := \exp\{-(1/p)\|u - v\|^2\}$. Both SVM classifiers were implemented by using the `svm` function in the `e1071` package (Meyer *et al.*, 2015). The GP classifier uses a radial basis function, with the hyperparameter chosen via the automatic method in the `gausspr` function in the `kernlab` package (Karatzoglou *et al.*, 2015). The tuning parameters for the other methods were chosen using the default settings in the corresponding R packages `PenLDA` (Witten, 2011), `NSC` (Hastie *et al.*, 2015) and `penalized` (Goeman *et al.*, 2015) namely sixfold, tenfold and fivefold cross-validation respectively. For the OTE and ES$k$nn methods we used the default settings in the R packages `OTE` (Khan *et al.*, 2015b) and `ESKNN` (Gul *et al.*, 2015).

6.1. Simulated examples

We present four simulation settings which were chosen to investigate the performance of the random-projection ensemble classifier in a wide variety of scenarios. In each of the examples below, we take $n \in \{50, 200, 1000\}$ and $p \in \{100, 1000\}$ and investigate two different values of the prior probability. We use Gaussian projections (see Section 5.1) and set $B_1 = 500$ and $B_2 = 50$ (see Section 5.3).

The risk estimates and standard errors for the $p = 100$ and $\pi_1 = 0.5$ case are shown in Tables 1 and 2 (the remaining results are given in the on-line supplementary material). These were calculated as follows: we set $n_{\text{test}} = 1000$ and $N_{\text{reps}} = 100$, and for $l = 1, \ldots, N_{\text{reps}}$ we generate a training set of size n and a test set of size n_{test} from the same distribution. Let $\hat{R}_l$ be the proportion of the test set that is classified incorrectly in the lth repeat of the experiment. The overall risk estimate presented is $\widehat{\text{risk}} := (1/N_{\text{reps}}) \Sigma_{l=1}^{N_{\text{reps}}} \hat{R}_l$. Note that

Table 1. Misclassification rates for models 1 and 2, with $p = 100$ and $\pi_1 = 0.5$

Method	*Results for model 1, Bayes risk 4.45*			*Results for model 2, Bayes risk 4.09*		
	n = 50	*n = 200*	*n = 1000*	*n = 50*	*n = 200*	*n = 1000*
RP-LDA$_2$	$49.34_{0.26}$	$48.10_{0.31}$	$44.14_{0.46}$	$8.34_{0.28}$	$\mathit{5.56}_{0.12}$	$\mathit{5.17}_{0.10}$
RP-LDA$_5$	$49.81_{0.24}$	$48.86_{0.30}$	$46.91_{0.40}$	$\mathit{8.17}_{0.27}$	$\mathit{5.64}_{0.13}$	$\mathit{5.14}_{0.10}$
RP-QDA$_2$	$44.18_{0.29}$	$29.38_{0.49}$	$10.57_{0.22}$	$8.40_{0.29}$	$\mathit{5.57}_{0.12}$	$\mathit{5.16}_{0.10}$
RP-QDA$_5$	$39.32_{0.33}$	$\mathit{22.32}_{0.32}$	$\mathit{8.75}_{0.15}$	$\mathit{8.06}_{0.25}$	$\mathit{5.58}_{0.12}$	$\mathit{5.09}_{0.10}$
RP-$k$nn$_2$	$46.10_{0.30}$	$36.18_{0.32}$	$19.42_{0.20}$	$8.94_{0.36}$	$\mathit{5.60}_{0.12}$	$5.20_{0.10}$
RP-$k$nn$_5$	$43.65_{0.30}$	$25.34_{0.35}$	$10.21_{0.16}$	$9.00_{0.33}$	$\mathit{5.68}_{0.12}$	$\mathit{5.13}_{0.10}$
LDA	—†	$49.60_{0.23}$	$49.91_{0.22}$	—†	$14.32_{0.22}$	$6.34_{0.11}$
QDA	—†	—†	$27.36_{0.23}$	—†	—†	$17.10_{0.20}$
$k$nn	$\mathit{34.66}_{0.35}$	$23.71_{0.31}$	$15.31_{0.17}$	$12.81_{0.28}$	$8.80_{0.15}$	$7.28_{0.13}$
RF	$49.72_{0.23}$	$48.33_{0.25}$	$43.28_{0.43}$	$11.11_{0.31}$	$6.80_{0.12}$	$6.07_{0.11}$
Radial SVM	$49.83_{0.22}$	$50.16_{0.22}$	$48.67_{0.22}$	$24.04_{1.47}$	$6.37_{0.14}$	$5.46_{0.10}$
Linear SVM	$50.02_{0.23}$	$49.55_{0.21}$	$50.04_{0.22}$	$9.41_{0.21}$	$8.96_{0.17}$	$7.76_{0.13}$
Radial GP	$48.18_{0.30}$	$42.76_{0.29}$	$26.60_{0.24}$	$14.09_{0.63}$	$5.84_{0.13}$	$\mathit{5.09}_{0.10}$
PenLDA	$49.95_{0.23}$	$49.79_{0.23}$	$50.05_{0.22}$	$11.11_{0.55}$	$6.72_{0.20}$	$5.79_{0.12}$
NSC	$49.74_{0.23}$	$49.69_{0.26}$	$49.55_{0.24}$	$12.61_{0.61}$	$7.27_{0.28}$	$5.82_{0.13}$
PenLog	$49.66_{0.35}$	$49.88_{0.24}$	$50.12_{0.21}$	$11.37_{0.22}$	$7.67_{0.14}$	$6.00_{0.11}$
SDR5-LDA	—†	$37.80_{0.48}$	$35.31_{0.30}$	—†	$15.07_{0.22}$	$6.47_{0.11}$
SDR5-$k$nn	—†	$32.22_{0.71}$	$21.83_{1.08}$	—†	$18.81_{0.29}$	$7.75_{0.12}$
OTE	$48.51_{0.33}$	$34.73_{1.23}$	$9.57_{0.66}$	$18.26_{0.47}$	$12.44_{0.26}$	$9.24_{0.15}$
ES$k$nn	$50.13_{0.23}$	$49.87_{0.22}$	$49.77_{0.21}$	$40.30_{0.71}$	$37.06_{0.63}$	$32.98_{0.58}$

†Not applicable.

Table 2. Misclassification rates for models 3 and 4, with $p=100$ and $\pi_1=0.5$

Method	*Results for model 3, Bayes risk 1.01*			*Results for model 4, Bayes risk 12.68*		
	n = 50	*n = 200*	*n = 1000*	*n = 50*	*n = 200*	*n = 1000*
RP-LDA$_2$	$45.11_{1.03}$	$44.05_{0.98}$	$39.22_{0.89}$	$38.06_{0.71}$	$38.45_{0.92}$	$40.48_{0.84}$
RP-LDA$_5$	$45.58_{0.60}$	$44.46_{0.58}$	$41.08_{0.56}$	$34.84_{0.63}$	$32.43_{0.75}$	$35.09_{0.89}$
RP-QDA$_2$	$11.41_{0.62}$	$4.83_{0.15}$	$3.85_{0.09}$	$42.12_{0.47}$	$41.99_{0.28}$	$42.37_{0.21}$
RP-QDA$_5$	$\mathit{9.71}_{0.52}$	$\mathit{4.23}_{0.14}$	$\mathit{3.29}_{0.08}$	$42.13_{0.35}$	$42.04_{0.27}$	$42.59_{0.21}$
RP-*k*nn$_2$	$20.69_{0.84}$	$6.86_{0.27}$	$4.73_{0.11}$	$30.85_{0.49}$	$24.07_{0.31}$	$20.76_{0.19}$
RP-*k*nn$_5$	$21.30_{0.54}$	$6.91_{0.18}$	$3.78_{0.10}$	$\mathit{29.85}_{0.46}$	$24.02_{0.30}$	$20.81_{0.21}$
LDA	—†	$46.22_{0.25}$	$41.74_{0.24}$	—†	$37.34_{0.29}$	$31.04_{0.26}$
QDA	—†	—†	$15.30_{0.21}$	—†	—†	$40.90_{0.21}$
*k*nn	$49.92_{0.24}$	$49.81_{0.22}$	$49.67_{0.22}$	$37.49_{0.63}$	$30.14_{0.34}$	$27.58_{0.25}$
RF	$44.79_{0.34}$	$23.38_{0.30}$	$7.72_{0.16}$	$30.97_{0.60}$	$\mathit{20.46}_{0.21}$	$\mathit{18.69}_{0.17}$
Radial SVM	$39.34_{1.47}$	$4.65_{0.13}$	$3.43_{0.09}$	$47.72_{0.40}$	$45.46_{0.51}$	$43.70_{0.72}$
Linear SVM	$46.57_{0.26}$	$46.17_{0.24}$	$41.67_{0.26}$	$36.79_{0.57}$	$34.21_{0.56}$	$31.87_{0.71}$
Radial GP	$48.87_{0.31}$	$45.47_{0.37}$	$36.18_{0.27}$	$38.39_{0.84}$	$26.63_{0.44}$	$22.77_{0.20}$
PenLDA	$46.04_{0.26}$	$44.48_{0.26}$	$41.71_{0.23}$	$45.64_{0.44}$	$45.22_{0.53}$	$45.39_{0.47}$
NSC	$47.47_{0.33}$	$45.99_{0.34}$	$42.31_{0.30}$	$46.34_{0.58}$	$44.69_{0.69}$	$45.72_{0.65}$
PenLog	$48.81_{0.29}$	$46.36_{0.28}$	$42.15_{0.24}$	—†	—†	—†
SDR5-LDA	—†	$46.27_{0.24}$	$42.09_{0.25}$	—†	$37.96_{0.29}$	$31.04_{0.27}$
SDR5-*k*nn	—†	$46.14_{0.27}$	$36.28_{0.24}$	—†	$39.70_{0.32}$	$29.31_{0.26}$
OTE	$46.74_{0.28}$	$30.62_{0.33}$	$11.43_{0.19}$	$32.24_{0.51}$	$23.37_{0.28}$	$19.59_{0.19}$
ES*k*nn	$48.66_{0.26}$	$46.59_{0.26}$	$45.17_{0.22}$	$46.15_{0.51}$	$44.03_{0.54}$	$43.77_{0.46}$

†Not applicable.

$$\mathbb{E}(\widehat{\text{risk}}) = \mathbb{E}\{R(C_n^{\text{RP}})\}$$

and

$$\begin{aligned}\text{var}(\widehat{\text{risk}}) &= \frac{1}{N_{\text{reps}}}\text{var}(\hat{R}_1)\\ &= \frac{1}{N_{\text{reps}}}\left\{\mathbb{E}\left(\frac{\mathbf{E}\{R(C_n^{\text{RP}})\}[1-\mathbf{E}\{R(C_n^{\text{RP}})\}]}{n_{\text{test}}}\right)+\text{var}[\mathbf{E}\{R(C_n^{\text{RP}})\}]\right\}.\end{aligned}$$

We therefore estimate the standard error in the tables below by

$$\hat{\sigma} := \frac{1}{N_{\text{reps}}^{1/2}}\left\{\frac{\widehat{\text{risk}}(1-\widehat{\text{risk}})}{n_{\text{test}}}+\frac{n_{\text{test}}-1}{n_{\text{test}}N_{\text{reps}}}\sum_{l=1}^{N_{\text{reps}}}(\hat{R}_l-\widehat{\text{risk}})^2\right\}^{1/2}.$$

The method with the smallest risk estimate in each column of the tables below is highlighted in italics; where applicable, we also highlight any method with a risk estimate within 1 standard error of the minimum.

6.1.1. *Sparse class boundaries: model 1*

Here, $X|\{Y=0\} \sim \frac{1}{2}N_p(\mu_0,\Sigma)+\frac{1}{2}N_p(-\mu_0,\Sigma)$, and $X|\{Y=1\} \sim \frac{1}{2}N_p(\mu_1,\Sigma)+\frac{1}{2}N_p(-\mu_1,\Sigma)$, where, for $p=100$, we set $\Sigma=I_{100\times100}$, $\mu_0=(2,-2,0,\ldots,0)^{\text{T}}$ and $\mu_1=(2,2,0,\ldots,0)^{\text{T}}$.

In model 1, assumption 3 holds with $d=2$; for example, we could take the rows of A^* to be the first two Euclidean basis vectors. We see that the random-projection ensemble classifier with the QDA base classifier performs very well here, as does the OTE method. Despite the fact that the regression function η depends on only the first two components in this example, the comparators that were designed for sparse problems do not perform well; in some cases they are no better than a random guess.

6.1.2. *Rotated sparse normal: model 2*

Here, $X|\{Y=0\} \sim N_p(\Omega_p\mu_0, \Omega_p\Sigma_0\Omega_p^{\mathrm{T}})$, and $X|\{Y=1\} \sim N_p(\Omega_p\mu_1, \Omega_p\Sigma_1\Omega_p^{\mathrm{T}})$, where Ω_p is a $p \times p$ rotation matrix that was sampled once according to Haar measure, and remained fixed thereafter, and we set $\mu_0 = (3,3,3,0,\ldots,0)^{\mathrm{T}}$ and $\mu_1 = (0,\ldots,0)^{\mathrm{T}}$. Moreover, Σ_0 and Σ_1 are block diagonal, with blocks $\Sigma_r^{(1)}$, and $\Sigma_r^{(2)}$, for $r=0,1$, where $\Sigma_0^{(1)}$ is a 3×3 matrix with diagonal entries equal to 2 and off-diagonal entries equal to $\frac{1}{2}$, and $\Sigma_1^{(1)} = \Sigma_0^{(1)} - I_{3\times 3}$. In both classes $\Sigma_r^{(2)}$ is a $(p-3)\times(p-3)$ matrix, with diagonal entries equal to 1 and off-diagonal entries equal to $\frac{1}{2}$.

In model 2, assumption 3 holds with $d=3$; for instance, A^* can be taken to be the first three rows of Ω_p^{T}. Perhaps surprisingly, whether we use too small a value of d (namely $d=2$), or a value that is too large ($d=5$), the random-projection ensemble methods still classify very well.

6.1.3. *Independent features: model 3*

Here, $P_0 = N_p(\mu, I_{p\times p})$, with $\mu = (1/\surd p)(1,\ldots,1,0,\ldots,0)^{\mathrm{T}}$, where μ has $p/2$ non-zero components, whereas P_1 is the distribution of p independent components, each with a standard Laplace distribution.

In model 3, the class boundaries are non-linear and, in fact, assumption 3 is not satisfied for any $d<p$. Nevertheless, in Table 2 we see that, where the LDA, QDA and $k$nn classifiers are tractable, they are outperformed by their random-projection ensemble counterparts and in fact the RP-QDA$_5$ classifier has the smallest misclassification rate among all methods implemented. Unsurprisingly, the methods that are designed for a linear Bayes decision boundary are not effective. The RP-QDA classifiers are especially accurate here; in particular, they can cope better with the non-linearity of the class boundaries than the RP-LDA classifiers.

6.1.4. *t-distributed features: model 4*

Here, $X|\{Y=r\} = \mu_r + Z_r/\surd(U_r/\nu_r)$, where $Z_r \sim N_p(0,\Sigma_r)$ independent of $U_r \sim \chi^2_{\nu_r}$, for $r=0,1$, i.e. P_r is the multivariate t-distribution centred at μ_r, with ν_r degrees of freedom and shape parameter Σ_r. We set $\mu_0 = (1,\ldots,1,0,\ldots,0)^{\mathrm{T}}$, where μ_0 has 10 non-zero components, $\mu_1 = 0$, $\nu_0 = 2$, $\nu_1 = 1$, $\Sigma_0 = (\Sigma_{j,k})$, where $\Sigma_{j,j} = 1$, $\Sigma_{j,k} = 0.5$ if $\max(j,k) \leqslant 10$ and $j \neq k$, $\Sigma_{j,k} = 0$ otherwise, and $\Sigma_1 = I_{p\times p}$.

Model 4 explores the effect of heavy tails and the presence of correlation between the features. Again, assumption 3 is not satisfied for any $d<p$. The RF, OTE and RP-$k$nn methods all perform very well here. The RP-LDA and RP-QDA classifiers are less good. This is partly because the class conditional distributions do not have finite second and first moments respectively and, as a result, the class mean and covariance matrix estimates are poor.

6.2. *Real data examples*

In this section, we compare the classifiers above on eight real data sets that are available from the University of California Irvine Machine Learning Repository. In each example, we first subsample the data to form a training set of size n and then use the remaining data (or, where available, take a subsample of size 1000 from it) to form the test set. As with the simulated examples, we set $B_1 = 500$ and $B_2 = 50$ and used Gaussian-distributed projections, and each experiment was repeated 100 times. Where appropriate, the tuning parameters were chosen via

Table 3. Misclassification rates for the eye state and ionosphere data sets

Method	*Results for eye state data*			*Results for ionosphere data*		
	$n=50$	$n=200$	$n=1000$	$n=50$	$n=100$	$n=200$
RP-LDA$_5$	$42.06_{0.38}$	$38.61_{0.29}$	$36.30_{0.21}$	$13.05_{0.38}$	$10.75_{0.25}$	$9.78_{0.26}$
RP-QDA$_5$	$\mathit{38.97}_{0.39}$	$32.44_{0.42}$	$30.91_{0.87}$	$\mathit{8.14}_{0.37}$	$\mathit{6.15}_{0.22}$	$\mathit{5.21}_{0.20}$
RP-*k*nn$_5$	$\mathit{39.37}_{0.39}$	$\mathit{26.91}_{0.27}$	$\mathit{13.54}_{0.19}$	$13.05_{0.46}$	$7.43_{0.25}$	$5.43_{0.19}$
LDA	$42.38_{0.40}$	$39.15_{0.30}$	$36.91_{0.23}$	$23.72_{0.40}$	$18.27_{0.28}$	$15.58_{0.31}$
QDA	$39.91_{0.35}$	$29.24_{0.40}$	$29.76_{1.07}$	—†	—†	$14.07_{0.34}$
*k*nn	$41.70_{0.40}$	$29.18_{0.27}$	$14.45_{0.16}$	$21.81_{0.73}$	$18.05_{0.46}$	$16.40_{0.35}$
RF	$\mathit{39.27}_{0.37}$	$29.04_{0.25}$	$17.63_{0.20}$	$10.52_{0.30}$	$7.54_{0.19}$	$6.48_{0.18}$
Radial SVM	$46.33_{0.49}$	$38.71_{0.46}$	$31.03_{0.68}$	$27.67_{1.15}$	$12.85_{0.91}$	$6.67_{0.22}$
Linear SVM	$42.38_{0.42}$	$39.55_{0.36}$	$38.58_{0.38}$	$19.41_{0.35}$	$17.05_{0.27}$	$15.48_{0.29}$
Radial GP	$40.73_{0.38}$	$32.22_{0.25}$	$21.66_{0.21}$	$22.29_{0.72}$	$17.81_{0.46}$	$14.52_{0.31}$
PenLDA	$44.37_{0.43}$	$42.50_{0.28}$	$41.86_{0.23}$	$21.20_{0.57}$	$19.83_{0.56}$	$19.81_{0.54}$
NSC	$44.73_{0.48}$	$42.37_{0.29}$	$42.27_{0.28}$	$22.62_{0.53}$	$19.11_{0.42}$	$17.52_{0.34}$
SDR5-LDA	$42.82_{0.40}$	$39.25_{0.29}$	$36.92_{0.23}$	$25.78_{0.52}$	$18.98_{0.30}$	$15.63_{0.30}$
SDR5-*k*nn	$42.43_{0.38}$	$34.13_{0.32}$	$25.31_{0.25}$	$30.61_{0.74}$	$17.53_{0.45}$	$10.12_{0.30}$
OTE	$40.10_{0.38}$	$29.92_{0.28}$	$18.73_{0.20}$	$14.38_{0.41}$	$9.80_{0.27}$	$7.33_{0.23}$
ES*k*nn	$45.62_{0.41}$	$43.06_{0.35}$	$39.37_{0.34}$	$27.81_{0.58}$	$23.23_{0.48}$	$20.05_{0.51}$

†Not applicable.

the methods that were described at the beginning of Section 6 for each of the 100 repeats of the experiment.

6.2.1. *Eye state detection*

The electroencephalogram eye state data set (`http://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`) consists of $p=14$ electroencephalogram measurements on 14980 observations. The task is to use the electroencephalogram reading to determine the state of the eye. There are 8256 observations for which the eye is open (class 0), and 6723 for which the eye is closed (class 1). Results are given in Table 3.

6.2.2. *Ionosphere data set*

The ionosphere data set (`http://archive.ics.uci.edu/ml/datasets/Ionosphere`) consists of $p=32$ high frequency antenna measurements on 351 observations. Observations are classified as good (class 0) or bad (class 1), depending on whether there is evidence for free electrons in the ionosphere or not. The class sizes are 225 (good) and 126 (bad). Results are given in Table 3.

6.2.3. *Down's syndrome diagnoses in mice*

The mice data set (`http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression`) consists of 570 healthy mice (class 0) and 507 mice with Down's syndrome (class 1). The task is to diagnose Down's syndrome on the basis of $p=77$ protein expression measurements. Results are given in Table 4.

6.2.4. *Hill–valley identification*

The hill–valley data set (`http://archive.ics.uci.edu/ml/datasets/Hill-Valley`)

Table 4. Misclassification rates for the mice and hill–valley data sets

Method	*Results for mice data*			*Results for hill–valley data*		
	$n=200$	$n=500$	$n=1000$	$n=100$	$n=200$	$n=500$
RP-LDA5	$25.17_{0.30}$	$23.56_{0.26}$	$23.35_{0.49}$	$\mathit{36.84}_{0.84}$	$\mathit{36.45}_{0.85}$	$\mathit{32.57}_{1.06}$
RP-QDA5	$18.24_{0.29}$	$16.05_{0.24}$	$15.45_{0.45}$	$44.43_{0.34}$	$43.56_{0.31}$	$41.10_{0.33}$
RP-*k*nn5	$11.24_{0.29}$	$\mathit{2.24}_{0.10}$	$\mathit{0.55}_{0.09}$	$49.08_{0.24}$	$47.27_{0.26}$	$36.39_{0.29}$
LDA	$\mathit{6.46}_{0.14}$	$3.38_{0.10}$	$2.17_{0.17}$	—†	$37.29_{0.48}$	$34.37_{0.36}$
*k*nn	$19.65_{0.26}$	$7.02_{0.17}$	$0.94_{0.13}$	$49.35_{0.24}$	$48.82_{0.21}$	$47.49_{0.24}$
RF	$7.94_{0.22}$	$2.41_{0.11}$	$\mathit{0.51}_{0.08}$	$48.32_{0.23}$	$47.23_{0.21}$	$44.11_{0.25}$
Radial SVM	$11.25_{0.29}$	$3.89_{0.13}$	$1.69_{0.16}$	$50.24_{0.19}$	$50.24_{0.19}$	$50.42_{0.21}$
Linear SVM	$\mathit{6.36}_{0.14}$	$3.64_{0.10}$	$2.51_{0.17}$	$48.56_{0.22}$	$47.03_{0.23}$	$44.84_{0.28}$
Radial GP	$21.22_{0.30}$	$13.78_{0.24}$	$8.66_{0.34}$	$48.33_{0.22}$	$47.24_{0.21}$	$45.11_{0.22}$
PenLDA	$26.10_{0.36}$	$24.07_{0.26}$	$23.91_{0.46}$	$49.59_{0.22}$	$49.73_{0.21}$	$49.55_{0.22}$
NSC	$30.30_{0.36}$	$28.06_{0.29}$	$28.47_{0.51}$	$49.87_{0.21}$	$49.91_{0.20}$	$49.92_{0.22}$
OTE	$11.83_{0.32}$	$6.26_{0.18}$	$3.26_{0.23}$	$48.33_{0.23}$	$47.18_{0.22}$	$44.20_{0.24}$
ES*k*nn	$39.03_{0.59}$	$34.33_{0.66}$	$31.65_{0.78}$	$49.31_{0.23}$	$48.90_{0.23}$	$48.03_{0.25}$

†Not applicable.

Table 5. Misclassification rates for the musk and cardiac arrhythmia data sets

Method	*Results for musk data*			*Results for arrhythmia data*		
	$n=100$	$n=200$	$n=500$	$n=50$	$n=100$	$n=200$
RP-LDA5	$14.63_{0.31}$	$12.18_{0.23}$	$10.15_{0.15}$	$33.24_{0.42}$	$30.19_{0.35}$	$27.49_{0.30}$
RP-QDA5	$12.08_{0.27}$	$9.92_{0.18}$	$8.64_{0.13}$	$\mathit{30.47}_{0.33}$	$28.28_{0.26}$	$26.31_{0.28}$
RP-*k*nn5	$\mathit{11.81}_{0.27}$	$\mathit{9.65}_{0.21}$	$8.04_{0.15}$	$33.49_{0.40}$	$30.18_{0.33}$	$27.09_{0.31}$
LDA	—†	$24.88_{0.42}$	$9.09_{0.15}$	—†	—†	—†
*k*nn	$14.68_{0.28}$	$11.75_{0.22}$	$8.20_{0.15}$	$40.64_{0.33}$	$38.94_{0.33}$	$35.76_{0.36}$
RF	$13.20_{0.20}$	$10.69_{0.18}$	$7.55_{0.13}$	$31.65_{0.39}$	$\mathit{26.72}_{0.29}$	$\mathit{22.40}_{0.31}$
Radial SVM	$15.25_{0.15}$	$15.21_{0.15}$	$15.00_{0.17}$	$48.39_{0.49}$	$47.24_{0.46}$	$46.85_{0.43}$
Linear SVM	$13.91_{0.25}$	$10.39_{0.18}$	$\mathit{7.41}_{0.12}$	$36.16_{0.47}$	$35.61_{0.39}$	$35.20_{0.35}$
Radial GP	$14.91_{0.16}$	$14.07_{0.20}$	$11.14_{0.19}$	$37.28_{0.42}$	$33.80_{0.40}$	$29.31_{0.35}$
PenLDA	$27.74_{0.58}$	$27.14_{0.54}$	$26.98_{0.31}$	—†	—†	—†
NSC	$15.32_{0.18}$	$15.22_{0.15}$	$15.20_{0.16}$	$34.98_{0.46}$	$33.00_{0.40}$	$31.08_{0.41}$
PenLog	$14.48_{0.28}$	$11.85_{0.21}$	—†	$34.92_{0.42}$	$30.48_{0.34}$	$26.12_{0.27}$
SDR5-LDA	—†	$25.12_{0.43}$	$9.08_{0.15}$	—†	—†	—†
SDR5-*k*nn	—†	$24.09_{0.62}$	$9.81_{0.16}$	—†	—†	—†
OTE	$13.90_{0.23}$	$11.04_{0.18}$	$8.05_{0.14}$	$33.90_{0.47}$	$27.83_{0.29}$	$23.75_{0.32}$
ES*k*nn	$19.55_{0.42}$	$18.09_{0.30}$	$16.07_{0.24}$	$45.86_{0.43}$	$45.62_{0.48}$	$43.41_{0.43}$

†Not applicable.

consists of 1212 observations of a terrain, each when plotted in sequence represents either a hill (class 0; size 600) or a valley (class 1; size 612). The task is to classify the terrain on the basis of a vector of dimension $p=100$. Results are given in Table 4.

6.2.5. *Musk identification*

The musk data set (`http://archive.ics.uci.edu/ml/datasets/Musk+\%28Version+2\%29`) consists of 1016 musk (class 0) and 5581 non-musk (class 1) molecules. The task is to classify a molecule on the basis of $p=166$ shape measurements. Results are given in Table 5.

Table 6. Misclassification rates for the activity recognition and Gisette data sets

Method	*Results for activity recognition data*			*Results for Gisette data*		
	$n=50$	$n=200$	$n=1000$	$n=50$	$n=200$	$n=1000$
RP-LDA$_5$	$0.18_{0.02}$	$0.10_{0.01}$	$0.01_{0.00}$	$15.75_{0.41}$	$10.58_{0.17}$	$9.39_{0.15}$
RP-QDA$_5$	$0.15_{0.02}$	$0.09_{0.01}$	$\mathit{0.00}_{0.00}$	$15.53_{0.40}$	$10.53_{0.19}$	$9.37_{0.16}$
RP-*k*nn$_5$	$0.21_{0.02}$	$0.11_{0.01}$	$0.01_{0.00}$	$15.95_{0.46}$	$11.09_{0.17}$	$9.57_{0.16}$
*k*nn	$0.26_{0.02}$	$0.13_{0.02}$	$0.02_{0.01}$	$18.41_{0.42}$	$10.44_{0.18}$	$5.64_{0.13}$
RF	$0.25_{0.02}$	$0.17_{0.02}$	$0.08_{0.01}$	$14.33_{0.47}$	$9.37_{0.15}$	$5.79_{0.12}$
Radial SVM	$1.58_{0.11}$	$0.89_{0.06}$	$0.18_{0.02}$	$50.03_{0.19}$	$50.41_{0.19}$	$50.79_{0.25}$
Linear SVM	$0.19_{0.02}$	$0.12_{0.01}$	$0.05_{0.01}$	$\mathit{11.92}_{0.27}$	$\mathit{6.82}_{0.11}$	$\mathit{4.45}_{0.11}$
Radial GP	$0.25_{0.02}$	$0.20_{0.02}$	$0.13_{0.01}$	$27.09_{1.32}$	$10.74_{0.21}$	$6.70_{0.13}$
PenLDA	$\mathit{0.11}_{0.02}$	$\mathit{0.04}_{0.01}$	$\mathit{0.00}_{0.00}$	—†	—†	—†
NSC	$0.29_{0.02}$	$0.24_{0.03}$	$0.06_{0.01}$	$15.72_{0.29}$	$13.63_{0.22}$	$12.83_{0.21}$
OTE	$0.61_{0.07}$	$0.38_{0.05}$	$0.09_{0.02}$	$14.18_{0.25}$	$9.69_{0.17}$	$6.24_{0.13}$
ES*k*nn	$1.74_{0.18}$	$0.88_{0.09}$	$0.41_{0.05}$	$45.76_{0.76}$	$44.81_{0.74}$	$44.45_{0.73}$

†Not applicable.

6.2.6. *Cardiac arrhythmia diagnoses*

The cardiac arrhythmia data set (`https://archive.ics.uci.edu/ml/datasets/Arrhythmia`) has one normal class of size 245 and 13 abnormal classes, which we combined to form a second class of size 206. We removed the nominal features and those with missing values, leaving $p=194$ electrocardiogram measurements. In this example, the PenLDA classifier is not applicable because some features have within-class standard deviation equal to 0. Results are given in Table 5.

6.2.7. *Human activity recognition*

The human activity recognition data set (`http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones`) consists of $p=561$ accelerometer measurements, recorded from a smartphone while a subject is performing an activity. We subsampled the data to include only the walking and laying activities. In the resulting data set, there are 1226 'walking' observations (class 0), and 1407 'laying' observations (class 1). Results are given in Table 6.

6.2.8. *Handwritten digits*

The Gisette data set (`https://archive.ics.uci.edu/ml/datasets/Gisette`) consists of 6000 observations of handwritten digits, namely 3000 '4's and 3000 '9's. Each observation represents the original 28×28 pixel image, with added noise variables resulting in a 5000-dimensional vector. We first subsampled 1500 of the 6000 observations, giving 760 '4's and 740 '9's—this data set was then kept fixed through the subsequent 100 repeats of the experiment. The observations are sparse with a large number of 0-entries. Results are given in Table 6.

6.3. *Conclusion of numerical study*

The numerical study above reveals the extremely encouraging finite sample performance that is achieved by the random-projection ensemble classifier. A random-projection ensemble method attains the lowest misclassification error in 23 of the 36 simulated and real data settings investigated, and in eight of the 13 remaining cases a random-projection ensemble method is in the top three of the classifiers considered. The flexibility that is offered by the random-projection ensemble classifier—in particular, the fact that any base classifier may be used—allows the practitioner to adapt the method to work well in a wide variety of problems.

Another key observation is that our assumption 3 is not necessary for the random-projection method to work well: in model 2, we achieve good results by using $d=2$, whereas assumption 3 holds only with a three- (or higher) dimensional projection. Moreover, even in situations where assumption 3 does not hold for any $d<p$, the random-projection method is still competitive; see in particular the results for model 3.

One example where the random-projection ensemble framework is not effective is for the Gisette data set. Here the data are very sparse; for each observation a large proportion of the features are exactly 0. Of course, applying a Gaussian or Haar random projection to an observation will remove the sparse structure. In this case, the practitioner may benefit by using an alternative distribution for the projections, such as axis-aligned projections (see the discussion in Section 7).

7. Discussion and extensions

We have introduced a general framework for high dimensional classification via the combination of the results of applying a base classifier on carefully selected low dimensional random projections of the data. One of its attractive features is its generality: the approach can be used in conjunction with any base classifier. Moreover, although we explored in detail one method for combining the random projections (partly because it facilitates rigorous statistical analysis), many other options are available here. For instance, instead of retaining only the projection within each block yielding the smallest estimate of test error, one might give weights to the different projections, where the weights decrease as the estimate of test error increases.

Many practical classification problems involve $K>2$ classes. The main issue in extending our methodology to such settings is the definition of C_n^{RP} analogous to expression (2). To outline one approach, let

$$\nu_{n,r}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_n^{\mathbf{A}_{b_1}}(x)=r\}}$$

for $r=0,1,\ldots,K-1$. Given $\alpha_0,\ldots,\alpha_{K-1}>0$ with $\Sigma_{r=0}^{K-1}\alpha_r=1$, we can then define

$$C_n^{\mathrm{RP}}(x) := \operatorname*{sarg\,max}_{r=0,\ldots,K-1} \{\alpha_r\,\nu_{n,r}(x)\},$$

where sarg max denotes the smallest element of the arg max in the case of a tie. The choice of $\alpha_0,\ldots,\alpha_{K-1}$ is analogous to the choice of α in the case $K=2$. It is therefore natural to seek to minimize the test error of the corresponding infinite simulation random-projection classifier as before.

In other situations, it may be advantageous to consider alternative types of projection, perhaps because of additional structure in the problem. One particularly interesting issue concerns ultrahigh dimensional settings, say p in the thousands. Here, it may be too time consuming to generate enough random projections to explore adequately the space $\mathcal{A}_{d\times p}$. As a mathematical quantification of this, the cardinality of an ϵ-net in the Euclidean norm of the surface of the Euclidean ball in $\mathbb{R}^p$ increases exponentially in p (e.g. Vershynin (2012)). In such challenging problems, one might restrict the projections $\mathbf{A}$ to be axis aligned, so that each row of $\mathbf{A}$ consists of a single non-zero component, equal to 1, and $p-1$ components equal to 0. There are then only $\binom{p}{d} \leqslant p^d/d!$ choices for the projections and, if d is small, it may be feasible even to carry out an exhaustive search. Of course, this approach loses one of the attractive features of our original proposal, namely the fact that it is equivariant to orthogonal transformations.

Nevertheless, corresponding theory can be obtained provided that the projection A^* in assumption 3 is axis aligned. This is a much stronger requirement, but it seems that imposing greater structure is inevitable to obtain good classification in such settings.

Our main focus in this work has been on the classification performance of the random-projection ensemble classifier, and not on the interpretability of the class assignments. However, the projections selected provide weights that give an indication of the relative importance of the different variables in the model. Another interesting direction, therefore, would be to understand the properties of the variable ranking that is induced by the random-projection ensemble classifier.

In conclusion, we believe that random projections offer many exciting possibilities for high dimensional data analysis. In a similar spirit to subsampling and bootstrap sampling, we can think of each random projection as a perturbation of our original data, and effects that are observed over many different perturbations are often the 'stable' effects that are sought by statisticians; see Meinshausen and Bühlmann (2010), and Shah and Samworth (2013) in the context of variable selection. Two of the key features that make them so attractive for classification problems are the ability to identify 'good' random projections from the data, and the fact that we can aggregate results from selected projections. We expect that these two properties will be important in identifying future application areas for related methodologies.

Acknowledgements

Both authors are supported by Engineering and Physical Sciences Research Council Fellowship EP/J017213/1; the second author is also supported by a Philip Leverhulme prize. The authors thank Rajen Shah, Ming Yuan and the reviewers for helpful comments.

Appendix A

A.1. Proof of theorem 1

Recall that the training data $\mathcal{T}_n=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ are fixed and the projections $\mathbf{A}_1,\mathbf{A}_2,\ldots,$ are independent and identically distributed in $\mathcal{A}$, independent of the pair (X,Y). The test error of the random-projection ensemble classifier has the representation

$$\begin{aligned}\mathbf{E}\{R(C_n^{\mathrm{RP}})\}&=\mathbf{E}\left\{\pi_0\int_{\mathbb{R}^p}\mathbb{1}_{\{C_n^{\mathrm{RP}}(x)=1\}}\mathrm{d}P_0(x)+\pi_1\int_{\mathbb{R}^p}\mathbb{1}_{\{C_n^{\mathrm{RP}}(x)=0\}}\mathrm{d}P_1(x)\right\}\\&=\mathbf{E}\left\{\pi_0\int_{\mathbb{R}^p}\mathbb{1}_{\{\nu_n(x)\geqslant\alpha\}}\mathrm{d}P_0(x)+\pi_1\int_{\mathbb{R}^p}\mathbb{1}_{\{\nu_n(x)<\alpha\}}\mathrm{d}P_1(x)\right\}\\&=\pi_0\int_{\mathbb{R}^p}\mathbf{P}\{\nu_n(x)\geqslant\alpha\}\,\mathrm{d}P_0(x)+\pi_1\int_{\mathbb{R}^p}\mathbf{P}\{\nu_n(x)<\alpha\}\,\mathrm{d}P_1(x),\end{aligned}$$

where $\nu_n(x)$ is defined in equation (1), and where the final equality follows by Fubini's theorem.

Let $U_{b_1}:=\mathbb{1}_{\{C_n^{\mathbf{A}_{b_1}}(X)=1\}}$, for $b_1=1,\ldots,B_1$. Then, conditionally on $\mu_n(X)=\theta\in[0,1]$, the random variables $U_1,\ldots,U_{B_1}$ are independent, each having a Bernoulli(θ) distribution. Recall that $G_{n,0}$ and $G_{n,1}$ are the distribution functions of $\mu_n(X)|\{Y=0\}$ and $\mu_n(X)|\{Y=1\}$ respectively. We can therefore write

$$\begin{aligned}\int_{\mathbb{R}^p}\mathbf{P}\{\nu_n(x)<\alpha\}\mathrm{d}P_1(x)&=\int_{[0,1]}\mathbb{P}\left\{\frac{1}{B_1}\sum_{b_1=1}^{B_1}U_{b_1}<\alpha|\hat{\mu}_n(X)=\theta\right\}\mathrm{d}G_{n,1}(\theta)\\&=\int_{[0,1]}\mathbb{P}(T<B_1\alpha)\mathrm{d}G_{n,1}(\theta),\end{aligned}$$

where here and throughout the proof T denotes a $\mathrm{Bin}(B_1,\theta)$ random variable. Similarly,

$$\int_{\mathbb{R}^p}\mathbf{P}\{\nu_n(x)\geqslant\alpha\}\mathrm{d}P_0(x)=1-\int_{[0,1]}\mathbb{P}(T<B_1\alpha)\mathrm{d}G_{n,0}(\theta).$$

It follows that

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} = \pi_0 + \int_{[0,1]} \mathbb{P}(T < B_1\alpha) \mathrm{d}G_n^{\circ}(\theta),$$

where $G_n^{\circ} := \pi_1 G_{n,1} - \pi_0 G_{n,0}$. Writing $g_n^{\circ} := \pi_1 g_{n,1} - \pi_0 g_{n,0}$, we now show that

$$\int_{[0,1]} \{\mathbb{P}(T < B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\} \mathrm{d}G_n^{\circ}(\theta) = \frac{1-\alpha-[[B_1\alpha]]}{B_1} g_n^{\circ}(\alpha) + \frac{\alpha(1-\alpha)}{2B_1} \dot{g}_n^{\circ}(\alpha) + o\left(\frac{1}{B_1}\right) \tag{18}$$

as $B_1 \to \infty$. Our proof involves a one-term Edgeworth expansion to the binomial distribution function in equation (18), where the error term is controlled uniformly in the parameter. The expansion relies on the following version of Esseen's smoothing lemma.

Theorem 4 (Esseen (1945), chapter 2, theorem 2b). Let c_1, C_1, $S > 0$, let $F: \mathbb{R} \to [0, \infty)$ be a non-decreasing function and let $G: \mathbb{R} \to \mathbb{R}$ be a function of bounded variation. Let $F^*(s) := \int_{-\infty}^{\infty} \exp(\mathrm{i}st) \mathrm{d}F(t)$ and $G^*(s) := \int_{-\infty}^{\infty} \exp(\mathrm{i}st) \mathrm{d}G(t)$ be the Fourier–Stieltjes transforms of F and G respectively. Suppose that

(a) $\lim_{t\to-\infty} F(t) = \lim_{t\to-\infty} G(t) = 0$ and $\lim_{t\to\infty} F(t) = \lim_{t\to\infty} G(t)$,
(b) $\int_{-\infty}^{\infty} |F(t) - G(t)| \mathrm{d}t < \infty$,
(c) the set of discontinuities of F and G is contained in $\{t_i : i \in \mathbb{Z}\}$, where (t_i) is a strictly increasing sequence with $\inf_i\{t_{i+1} - t_i\} \geqslant c_1$ (moreover F is constant on the intervals $[t_i, t_{i+1})$ for all $i \in \mathbb{Z}$) and
(d) $|\dot{G}(t)| \leqslant C_1$ for all $t \notin \{t_i : i \in \mathbb{Z}\}$.

Then there are constants $c_2, C_2 > 0$ such that

$$\sup_{t\in\mathbb{R}} |F(t) - G(t)| \leqslant \frac{1}{\pi} \int_{-S}^{S} \left| \frac{F^*(s) - G^*(s)}{s} \right| \mathrm{d}s + \frac{C_1 C_2}{S},$$

provided that $Sc_1 \geqslant c_2$.

Let $\sigma^2 := \theta(1-\theta)$, and let Φ and ϕ denote the standard normal distribution and density functions respectively. Moreover, for $t \in \mathbb{R}$, let

$$p(t) = p(t, \theta) := \frac{(1-t^2)(1-2\theta)}{6\sigma}$$

and

$$q(t) = q(t, B_1, \theta) := \frac{\frac{1}{2} - [[B_1\theta + B_1^{1/2}\sigma t]]}{\sigma}.$$

In proposition 3 below we apply theorem 4 to the functions

$$F_{B_1}(t) = F_{B_1}(t, \theta) := \mathbb{P}\left(\frac{T - B_1\theta}{B_1^{1/2}\sigma} < t \right) \tag{19}$$

and

$$G_{B_1}(t) = G_{B_1}(t, \theta) := \Phi(t) + \phi(t) \frac{p(t, \theta) + q(t, B_1, \theta)}{B_1^{1/2}}. \tag{20}$$

Proposition 3. Let F_{B_1} and G_{B_1} be as in expressions (19) and (20). There is a constant $C > 0$ such that, for all $B_1 \in \mathbb{N}$,

$$\sup_{\theta\in(0,1)} \sup_{t\in\mathbb{R}} \sigma^3 |F_{B_1}(t, \theta) - G_{B_1}(t, \theta)| \leqslant \frac{C}{B_1}.$$

Proposition 3, whose proof is given after the proof of proposition 2, bounds uniformly in θ the error in the one-term Edgeworth expansion G_{B_1} of the distribution function F_{B_1}. Returning to the proof of theorem 1, we shall argue that the dominant contribution to the integral in equation (18) arises from the interval $(\max\{0, \alpha-\epsilon_1\}, \min\{\alpha+\epsilon_1, 1\})$, where $\epsilon_1 := B_1^{-1/2}\log(B_1)$. For the remainder of the proof we assume that B_1 is sufficiently large that $[\alpha-\epsilon_1, \alpha+\epsilon_1] \subseteq (0,1)$.

For the region $|\theta-\alpha| \geqslant \epsilon_1$, by Hoeffding's inequality, we have that

$$\sup_{|\theta-\alpha|\geqslant\epsilon_1} |\mathbb{P}(T<B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}| \leqslant \sup_{|\theta-\alpha|\geqslant\epsilon_1} \exp\{-2B_1(\theta-\alpha)^2\} \leqslant \exp\{-2\log^2(B_1)\} = O(B_1^{-M}),$$

for each $M>0$, as $B_1\rightarrow\infty$. Writing $I := [\alpha-\epsilon_1, \alpha+\epsilon_1]$, it follows that

$$\int_{[0,1]} \{\mathbb{P}(T<B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\} \mathrm{d}G_n^\circ(\theta) = \int_I \{\mathbb{P}(T<B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\} \mathrm{d}G_n^\circ(\theta) + O(B_1^{-M}), \tag{21}$$

for each $M>0$, as $B_1\rightarrow\infty$.

For the region $|\theta-\alpha|<\epsilon_1$, by proposition 3, there exists $C'>0$ such that, for all B_1 sufficiently large,

$$\sup_{|\theta-\alpha|<\epsilon_1} \left| \mathbb{P}(T<B_1\alpha) - \Phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} - \frac{1}{B_1^{1/2}}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} r\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} \right| \leqslant \frac{C'}{B_1},$$

where $r(t) := p(t)+q(t)$. Hence, using the fact that, for large B_1, $\sup_{|\theta-\alpha|<\epsilon_1} |g_n^\circ(\theta)| \leqslant |g_n^\circ(\alpha)| + 1 < \infty$ under assumption 1, we have

$$\begin{aligned} \int_I \{\mathbb{P}(T<B_1\alpha) - \mathbb{1}_{\{\theta<\alpha\}}\} \mathrm{d}G_n^\circ(\theta) &= \int_I \left[\Phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} - \mathbb{1}_{\{\theta<\alpha\}}\right] \mathrm{d}G_n^\circ(\theta) \\ &\quad + \frac{1}{B_1^{1/2}} \int_I \phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} r\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} \mathrm{d}G_n^\circ(\theta) + o\left(\frac{1}{B_1}\right), \end{aligned} \tag{22}$$

as $B_1\rightarrow\infty$. To aid exposition, we shall henceforth concentrate on the dominant terms in our expansions, denoting the remainder terms as $R_1, R_2, \ldots$. These remainders are then controlled at the end of the argument. For the first term in equation (22), we write

$$\begin{aligned} \int_I \left[\Phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} - \mathbb{1}_{\{\theta<\alpha\}}\right] \mathrm{d}G_n^\circ(\theta) &= \int_I \left(\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] - \mathbb{1}_{\{\theta<\alpha\}}\right) \mathrm{d}G_n^\circ(\theta) \\ &\quad + \frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}} \int_I (\alpha-\theta)^2 \phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] \mathrm{d}G_n^\circ(\theta) + R_1. \end{aligned} \tag{23}$$

Now, for the first term in equation (23),

$$\begin{aligned} \int_I \left(\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] - \mathbb{1}_{\{\theta<\alpha\}}\right) \mathrm{d}G_n^\circ(\theta) &= \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left(\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] - \mathbb{1}_{\{\theta<\alpha\}}\right) \{g_n^\circ(\alpha) + (\theta-\alpha)\,\dot{g}_n^\circ(\alpha)\} \mathrm{d}\theta + R_2 \\ &= \frac{\surd\{\alpha(1-\alpha)\}}{B_1^{1/2}} \int_{-\infty}^{\infty} \{\Phi(-u) - \mathbb{1}_{\{u<0\}}\} \left[g_n^\circ(\alpha) + \frac{\surd\{\alpha(1-\alpha)\}}{B_1^{1/2}} u\,\dot{g}_n^\circ(\alpha)\right] \mathrm{d}u \\ &\quad + R_2 + R_3 \\ &= \frac{\alpha(1-\alpha)}{2B_1} \dot{g}_n^\circ(\alpha) + R_2 + R_3. \end{aligned} \tag{24}$$

For the second term in equation (23), write

$$\begin{aligned} &\frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}} \int_I (\alpha-\theta)^2 \phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] \mathrm{d}G_n^\circ(\theta) \\ &\quad = \frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}} g_n^\circ(\alpha) \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} (\alpha-\theta)^2 \phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right] \mathrm{d}\theta + R_4 \\ &\quad = \frac{\frac{1}{2}-\alpha}{B_1} g_n^\circ(\alpha) \int_{-\infty}^{\infty} u^2 \phi(-u) \mathrm{d}u + R_4 + R_5 = \frac{\frac{1}{2}-\alpha}{B_1} g_n^\circ(\alpha) + R_4 + R_5. \end{aligned} \tag{25}$$

Returning to the second term in equation (22), observe that

$$\begin{aligned}\frac{1}{B_1^{1/2}}\int_I \phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\} r\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}\mathrm{d}G_n^\circ(\theta)&\\ =\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1^{1/2}}\int_I \frac{1}{\sigma}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}\mathrm{d}G_n^\circ(\theta)&\\ +\frac{1}{6B_1^{1/2}}\int_I \frac{1-2\theta}{\sigma}\left\{1-\frac{B_1(\alpha-\theta)^2}{\sigma^2}\right\}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}\mathrm{d}G_n^\circ(\theta)&\\ =\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1^{1/2}}\int_I \frac{1}{\sigma}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}\mathrm{d}G_n^\circ(\theta)+R_6&\\ =\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1^{1/2}\surd\{\alpha(1-\alpha)\}}g_n^\circ(\alpha)\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]\mathrm{d}\theta+R_6+R_7&\\ =\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1}g_n^\circ(\alpha)+R_6+R_7+R_8.&\end{aligned}\tag{26}$$

The claim (18) will now follow from equations (21)–(26), once we have shown that

$$\sum_{j=1}^{8}|R_j|=o(B_1^{-1})\tag{27}$$

as $B_1\to\infty$.

(a) *To bound* R_1, for $\zeta\in(0,1)$, let

$$h_\theta(\zeta):=\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\zeta(1-\zeta)\}}\right].$$

Observe that, by a Taylor series expansion about $\zeta=\alpha$, there exists $B_0\in\mathbb{N}$, such that, for all $B_1>B_0$ and all $\theta,\zeta\in(\alpha-\epsilon_1,\alpha+\epsilon_1)$,

$$\begin{aligned}\left|\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\zeta(1-\zeta)\}}\right]-\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]+(\zeta-\alpha)\frac{(1-2\alpha)B_1^{1/2}(\alpha-\theta)}{2\{\alpha(1-\alpha)\}^{3/2}}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]\right|&\\ =|h_\theta(\zeta)-h_\theta(\alpha)-(\zeta-\alpha)\dot{h}_\theta(\alpha)|&\\ \leqslant\frac{(\zeta-\alpha)^2}{2}\sup_{\zeta'\in[\alpha-\zeta,\alpha+\zeta]}|\ddot{h}_\theta(\zeta')|&\\ \leqslant(\zeta-\alpha)^2\frac{\log^3(B_1)}{2\surd(2\pi)\{\alpha(1-\alpha)\}^{7/2}}.&\end{aligned}$$

Using this bound with $\zeta=\theta$, we deduce that, for all B_1 sufficiently large,

$$\begin{aligned}|R_1|=\left|\int_I\left(\Phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}-\Phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]-\frac{(1-2\alpha)B_1^{1/2}(\alpha-\theta)^2}{2\{\alpha(1-\alpha)\}^{3/2}}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]\right)\mathrm{d}G_n^\circ(\theta)\right|&\\ \leqslant\frac{\log^3(B_1)}{2\surd(2\pi)\{\alpha(1-\alpha)\}^{7/2}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}(\theta-\alpha)^2|g_n^\circ(\theta)|\mathrm{d}\theta&\\ \leqslant\frac{\log^6(B_1)}{3\surd(2\pi)B_1^{3/2}\{\alpha(1-\alpha)\}^{7/2}}\sup_{|\theta-\alpha|\leqslant\epsilon_1}|g_n^\circ(\theta)|=o\left(\frac{1}{B_1}\right)&\end{aligned}$$

as $B_1\to\infty$.

(b) *To bound*R_2, since g_n° is differentiable at α, given $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$|g_n^\circ(\theta) - g_n^\circ(\alpha) - (\theta - \alpha)\dot{g}_n^\circ(\alpha)| < \epsilon|\theta - \alpha|,$$

for all $|\theta - \alpha| < \delta_\epsilon$. It follows that, for all B_1 sufficiently large,

$$\begin{aligned}
|R_2| = &\left| \int_I \left(\Phi\left[\frac{B_1^{1/2}(\alpha - \theta)}{\surd\{\alpha(1-\alpha)\}} \right] - \mathbb{1}_{\{\theta<\alpha\}} \right) \mathrm{d}G_n^\circ(\theta) \right. \\
&\left. - \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left(\Phi\left[\frac{B_1^{1/2}(\alpha - \theta)}{\surd\{\alpha(1-\alpha)\}} \right] - \mathbb{1}_{\{\theta<\alpha\}} \right) \{g_n^\circ(\alpha) + (\theta - \alpha)\dot{g}_n^\circ(\alpha)\} \mathrm{d}\theta \right| \\
\leqslant &\, \epsilon \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left| \Phi\left[\frac{B_1^{1/2}(\alpha - \theta)}{\surd\{\alpha(1-\alpha)\}} \right] - \mathbb{1}_{\{\theta<\alpha\}} \right| |\theta - \alpha| \mathrm{d}\theta \\
\leqslant &\, \frac{\epsilon\alpha(1-\alpha)}{B_1} \int_{-\log(B_1)/\surd\{\alpha(1-\alpha)\}}^{\log(B_1)/\surd\{\alpha(1-\alpha)\}} |\Phi(-u) - \mathbb{1}_{\{u<0\}}||u| \mathrm{d}u \\
\leqslant &\, \frac{2\epsilon\alpha(1-\alpha)}{B_1} \int_0^\infty u\,\Phi(-u)\mathrm{d}u = \frac{\epsilon\alpha(1-\alpha)}{2B_1}.
\end{aligned}$$

We deduce that $|R_2| = o(B_1^{-1})$ as $B_1 \to \infty$.

(c) *To bound* R_3, for large B_1, we have

$$\begin{aligned}
|R_3| = &\left| \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} \left(\Phi\left[\frac{B_1^{1/2}(\alpha - \theta)}{\surd\{\alpha(1-\alpha)\}} \right] - \mathbb{1}_{\{\theta<\alpha\}} \right) \{g_n^\circ(\alpha) + (\theta - \alpha)\dot{g}_n^\circ(\alpha)\} \mathrm{d}\theta \right. \\
&\left. - \frac{\surd\{\alpha(1-\alpha)\}}{B_1^{1/2}} \int_{-\infty}^{\infty} \{\Phi(-u) - \mathbb{1}_{\{u<0\}}\} \left[g_n^\circ(\alpha) + \frac{\surd\{\alpha(1-\alpha)\}}{B_1^{1/2}} u\, \dot{g}_n^\circ(\alpha) \right] \mathrm{d}u \right| \\
= &\, \frac{2\alpha(1-\alpha)}{B_1} |\dot{g}_n^\circ(\alpha)| \int_{\epsilon_1 B_1^{1/2}/\{\alpha(1-\alpha)\}^{1/2}}^{\infty} u\Phi(-u)\mathrm{d}u \\
\leqslant &\, \frac{2\{\alpha(1-\alpha)\}^{3/2}}{B_1 \log(B_1)} |\dot{g}_n^\circ(\alpha)| \int_0^\infty u^2\Phi(-u)\mathrm{d}u = \frac{2\surd 2\{\alpha(1-\alpha)\}^{3/2}}{3\surd\pi B_1 \log(B_1)} |\dot{g}_n^\circ(\alpha)| = o(B_1^{-1})
\end{aligned}$$

as $B_1 \to \infty$.

(d) *To bound* R_4, since g_n° is continuous at α, given $\epsilon > 0$, there exists $B_0' \in \mathbb{N}$ such that, for all $B_1 > B_0'$,

$$\sup_{|\theta-\alpha|\leqslant\epsilon_1} |g_n^\circ(\theta) - g_n^\circ(\alpha)| < \epsilon. \qquad (28)$$

Hence, given $\epsilon > 0$, for all $B_1 > B_0'$

$$\begin{aligned}
|R_4| = &\left| \frac{(1-2\alpha)B_1^{1/2}}{2\{\alpha(1-\alpha)\}^{3/2}} \int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1} (\alpha - \theta)^2 \phi\left[\frac{B_1^{1/2}(\alpha - \theta)}{\surd\{\alpha(1-\alpha)\}} \right] \{g_n^\circ(\theta) - g_n^\circ(\alpha)\} \mathrm{d}\theta \right| \\
\leqslant &\, \frac{\epsilon|1-2\alpha|}{2B_1} \int_{-\infty}^{\infty} u^2\phi(-u)\mathrm{d}u = \frac{\epsilon|1-2\alpha|}{2B_1}.
\end{aligned}$$

(e) *To bound*R_5, for all B_1 sufficiently large,

$$\begin{aligned}
|R_5| = &\, \frac{|1-2\alpha|}{B_1} |g_n^\circ(\alpha)| \int_{\log(B_1)/\surd\{\alpha(1-\alpha)\}}^{\infty} u^2\phi(-u)\mathrm{d}u \\
\leqslant &\, \frac{\surd\{\alpha(1-\alpha)\}}{B_1 \log(B_1)} |g_n^\circ(\alpha)| \int_0^\infty u^3\phi(-u)\mathrm{d}u = \frac{\surd\{2\alpha(1-\alpha)\}}{\surd\pi B_1 \log(B_1)} |g_n^\circ(\alpha)| = o\left(\frac{1}{B_1}\right)
\end{aligned}$$

as $B_1 \to \infty$.

(f) *To bound*R_6, we write $R_6=R_{61}+R_{62}$, where

$$R_{61}:=\frac{1-2\alpha}{6B_1^{1/2}\sqrt{\{\alpha(1-\alpha)\}}}\int_I\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\mathrm{d}G_n^\circ(\theta)$$

and

$$\begin{aligned}R_{62}:=&\frac{1}{6B_1^{1/2}}\int_I\frac{1-2\theta}{\sigma}\left\{1-\frac{B_1(\alpha-\theta)^2}{\sigma^2}\right\}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}\mathrm{d}G_n^\circ(\theta)\\&-\frac{(1-2\alpha)}{6B_1^{1/2}\sqrt{\{\alpha(1-\alpha)\}}}\int_I\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\mathrm{d}G_n^\circ(\theta).\end{aligned}$$

By the bound (28), it follows that, for $B_1>B_0'$ sufficiently large,

$$\begin{aligned}|R_{61}|\leqslant&\frac{|1-2\alpha|}{6B_1^{1/2}\sqrt{\{\alpha(1-\alpha)\}}}|g_n^\circ(\alpha)|\left|\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\mathrm{d}\theta\right|\\&+\epsilon\frac{|1-2\alpha|}{6B_1^{1/2}\sqrt{\{\alpha(1-\alpha)\}}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\left|1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right|\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\mathrm{d}\theta.\\\leqslant&\frac{|1-2\alpha|}{6B_1}|g_n^\circ(\alpha)|\left|\int_{-\log(B_1)/\sqrt{\{\alpha(1-\alpha)\}}}^{\log(B_1)/\sqrt{\{\alpha(1-\alpha)\}}}(1-u^2)\phi(-u)\mathrm{d}u\right|\\&+\epsilon\frac{|1-2\alpha|}{6B_1}\int_{-\infty}^{\infty}(1+u^2)\phi(-u)\mathrm{d}u\leqslant\frac{\epsilon}{B_1}.\end{aligned}$$

We deduce that $R_{61}=o(B_1^{-1})$ as $B_1\rightarrow\infty$.

To control R_{62}, by the mean value theorem, we have that, for all B_1 sufficiently large and all $\zeta\in[\alpha-\epsilon_1,\alpha+\epsilon_1]$,

$$\begin{aligned}\sup_{|\theta-\alpha|<\epsilon_1}&\left|\frac{1-2\zeta}{\sqrt{\{\zeta(1-\zeta)\}}}\left\{1-\frac{B_1(\alpha-\theta)^2}{\zeta(1-\zeta)}\right\}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\zeta(1-\zeta)\}}}\right]\right.\\&\left.-\frac{(1-2\alpha)}{\sqrt{\{\alpha(1-\alpha)\}}}\left\{1-\frac{B_1(\alpha-\theta)^2}{\alpha(1-\alpha)}\right\}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\right|\\&\leqslant\frac{\log^4(B_1)}{\sqrt{(2\pi)}\{\alpha(1-\alpha)\}^{7/2}}|\zeta-\alpha|.\end{aligned}$$

Thus, for large B_1,

$$\begin{aligned}|R_{62}|&\leqslant\frac{\log^4(B_1)}{6\sqrt{(2\pi)}B_1^{1/2}\{\alpha(1-\alpha)\}^{7/2}}\sup_{|\theta-\alpha|\leqslant\epsilon_1}|g_n^\circ(\theta)|\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}|\theta-\alpha|\mathrm{d}\theta\\&\leqslant\frac{\log^6[B_1\{1+|g_n^\circ(\alpha)|\}]}{6\sqrt{(2\pi)}B_1^{3/2}\{\alpha(1-\alpha)\}^{7/2}}=o\left(\frac{1}{B_1}\right).\end{aligned}$$

We deduce that $|R_6|=o(B_1^{-1})$ as $B_1\rightarrow\infty$.

(g) *To bound* R_7, write $R_7=R_{71}+R_{72}$, where

$$R_{71}:=\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1^{1/2}\sqrt{\{\alpha(1-\alpha)\}}}\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\{g_n^\circ(\theta)-g_n^\circ(\alpha)\}\mathrm{d}\theta$$

and

$$R_{72}:=\frac{\frac{1}{2}-[[B_1\alpha]]}{B_1^{1/2}}\int_I\left(\frac{1}{\sigma}\phi\left\{\frac{B_1^{1/2}(\alpha-\theta)}{\sigma}\right\}-\frac{1}{\sqrt{\{\alpha(1-\alpha)\}}}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\sqrt{\{\alpha(1-\alpha)\}}}\right]\right)\mathrm{d}G_n^\circ(\theta).$$

By the bound (28), given $\epsilon > 0$, for all B_1 sufficiently large,

$$|R_{71}| \leqslant \frac{\epsilon}{2B_1^{1/2}\surd\{\alpha(1-\alpha)\}}\int_{-\infty}^{\infty}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]\mathrm{d}\theta = \frac{\epsilon}{2B_1}.$$

Moreover, by the mean value theorem, for all B_1 sufficiently large and all $|\zeta-\alpha| \leqslant \epsilon_1$,

$$\sup_{|\theta-\alpha|<\epsilon_1}\left|\frac{1}{\surd\{\zeta(1-\zeta)\}}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\zeta(1-\zeta)\}}\right] - \frac{1}{\surd\{\alpha(1-\alpha)\}}\phi\left[\frac{B_1^{1/2}(\alpha-\theta)}{\surd\{\alpha(1-\alpha)\}}\right]\right|$$
$$\leqslant \frac{\log^2(B_1)}{\surd(2\pi)\{\alpha(1-\alpha)\}^{5/2}}|\zeta-\alpha|.$$

It follows that, for all B_1 sufficiently large,

$$\begin{aligned}|R_{72}| &\leqslant \frac{\log^2(B_1)}{2\surd(2\pi)B_1^{1/2}\{\alpha(1-\alpha)\}^{5/2}}\sup_{|\theta-\alpha|\leqslant\epsilon_1}|g_n^\circ(\theta)|\int_{\alpha-\epsilon_1}^{\alpha+\epsilon_1}|\theta-\alpha|\mathrm{d}\theta\\ &\leqslant \frac{\log^4[B_1\{1+|g_n^\circ(\alpha)|\}]}{2\surd(2\pi)B_1^{3/2}\{\alpha(1-\alpha)\}^{5/2}}.\end{aligned}$$

We deduce that $|R_7| = o(B_1^{-1})$ as $B_1 \to \infty$.

(h) *To bound* R_8, we have

$$|R_8| = \frac{2(\frac{1}{2}-[[B_1\alpha]])}{B_1}|g_n^\circ(\alpha)|\int\limits_{\epsilon_1 B_1^{1/2}/\{\alpha(1-\alpha)\}^{1/2}}^{\infty}\phi(-u)\mathrm{d}u = o\left(\frac{1}{B_1}\right)$$

as $B_1 \to \infty$.

We have now established claim (27), and the result follows.

A.2. Proof of theorem 2

In the case where $B_1 < \infty$, we have

$$\begin{aligned}R(C_n^{\mathrm{RP}}) - R(C^{\mathrm{Bayes}}) &= \int_{\mathbb{R}^p}[\eta(x)(\mathbb{1}_{\{C_n^{\mathrm{RP}}(x)=0\}} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(x)=0\}}) + \{1-\eta(x)\}(\mathbb{1}_{\{C_n^{\mathrm{RP}}(x)=1\}} - \mathbb{1}_{\{C^{\mathrm{Bayes}}(x)=1\}})]\mathrm{d}P_X(x)\\ &= \int_{\mathbb{R}^p}\{|2\eta(x)-1||\mathbb{1}_{\{\nu_n(x)<\alpha\}} - \mathbb{1}_{\{\eta(x)<1/2\}}|\}\mathrm{d}P_X(x)\\ &= \int_{\mathbb{R}^p}\{|2\eta(x)-1|\mathbb{1}_{\{\nu_n(x)\geqslant\alpha\}}\mathbb{1}_{\{\eta(x)<1/2\}} + |2\eta(x)-1|\mathbb{1}_{\{\nu_n(x)<\alpha\}}\mathbb{1}_{\{\eta(x)\geqslant1/2\}}\}\mathrm{d}P_X(x)\\ &\leqslant \int_{\mathbb{R}^p}\left[\frac{1}{\alpha}|2\eta(x)-1|\nu_n(x)\mathbb{1}_{\{\eta(x)<1/2\}} + \frac{1}{1-\alpha}|2\eta(x)-1|\{1-\nu_n(x)\}\mathbb{1}_{\{\eta(x)\geqslant1/2\}}\right]\mathrm{d}P_X(x).\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}}) &\leqslant \mathbf{E}\left\{\int_{\mathbb{R}^p}\frac{1}{\alpha}|2\eta(x)-1|\mathbb{1}_{\{C_n^{\mathbf{A}_1}(x)=1\}}\mathbb{1}_{\{\eta(x)<1/2\}}\right.\\ &\qquad\left. + \frac{1}{1-\alpha}|2\eta(x)-1|\mathbb{1}_{\{C_n^{\mathbf{A}_1}(x)=0\}}\mathbb{1}_{\{\eta(x)\geqslant1/2\}}\mathrm{d}P_X(x)\right\}\\ &\leqslant \frac{1}{\min(\alpha,1-\alpha)}\mathbf{E}\left\{\int_{\mathbb{R}^p}|2\eta(x)-1||\mathbb{1}_{\{C_n^{\mathbf{A}_1}(x)=0\}} - \mathbb{1}_{\{\eta(x)<1/2\}}|\mathrm{d}P_X(x)\right\}\\ &= \frac{1}{\min(\alpha,1-\alpha)}[\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\mathrm{Bayes}})],\end{aligned}$$

as required. When $B_1 = \infty$, we replace both occurrences of $R(C_n^{\mathrm{RP}})$ with $R(C_n^{\mathrm{RP}*})$ and the argument goes through in almost identical fashion after changing ν_n to μ_n.

A.3. Proof of theorem 3

First write

$$\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\text{Bayes}}) = \mathbf{E}(R_n^{\mathbf{A}_1}) - R(C^{\text{Bayes}}) + \epsilon_n.$$

Using assumption 2, we have that

$$\begin{aligned}\mathbf{E}(R_n^{\mathbf{A}_1}) &= \mathbf{E}(R_n^{\mathbf{A}_1} \mathbb{1}_{\{R_n^{\mathbf{A}_1} \leqslant R_n^* + |\epsilon_n|\}}) + \mathbf{E}(R_n^{\mathbf{A}_1} \mathbb{1}_{\{R_n^{\mathbf{A}_1} > R_n^* + |\epsilon_n|\}}) \\ &\leqslant R_n^* + |\epsilon_n| + \mathbf{P}(R_n^{\mathbf{A}_1} > R_n^* + |\epsilon_n|) \\ &= R_n^* + |\epsilon_n| + \mathbf{P}(R_n^{\mathbf{A}_{1,1}} > R_n^* + |\epsilon_n|)^{B_2} \\ &\leqslant R_n^* + |\epsilon_n| + (1-\beta)^{B_2}.\end{aligned}$$

But, for any $A \in \mathcal{A}$ and by definition of R_n^* and ϵ_n^A, we have $R_n^* \leqslant R_n^A = R(C_n^A) - \epsilon_n^A$. It therefore follows by theorem 2 that

$$\begin{aligned}\mathbf{E}\{R(C_n^{\text{RP}})\} - R(C^{\text{Bayes}}) &\leqslant \frac{1}{\min(\alpha, 1-\alpha)} [\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\text{Bayes}})] \\ &\leqslant \frac{R(C_n^A) - R(C^{\text{Bayes}})}{\min(\alpha, 1-\alpha)} + \frac{2|\epsilon_n| - \epsilon_n^A}{\min(\alpha, 1-\alpha)} + \frac{(1-\beta)^{B_2}}{\min(\alpha, 1-\alpha)},\end{aligned}$$

as required.

A.4. Proof of proposition 1

For a Borel set $C \subseteq \mathbb{R}^d$, let $P_{A^*X}(C) := \int_{\{x: A^*x \in C\}} \mathrm{d}P_X(x)$, so that P_{A^*X} is the marginal distribution of A^*X. Further, for $z \in \mathbb{R}^d$, write $P_{X|A^*X=z}$ for the conditional distribution of X given $A^*X = z$. If Y is independent of X given A^*X, and if B is a Borel subset of $\mathbb{R}^p$, then

$$\begin{aligned}\int_B \eta^{A^*}(A^*x) \mathrm{d}P_X(x) &= \int_{\mathbb{R}^d} \int_{B \cap \{w: A^*w = z\}} \eta^{A^*}(A^*w) \mathrm{d}P_{X|A^*X=z}(w) \mathrm{d}P_{A^*X}(z) \\ &= \int_{\mathbb{R}^d} \eta^{A^*}(z)\, \mathbb{P}(X \in B | A^*X = z) \mathrm{d}P_{A^*X}(z) \\ &= \int_{\mathbb{R}^d} \mathbb{P}(Y=1, X \in B | A^*X = z) \mathrm{d}P_{A^*X}(z) \\ &= \mathbb{P}(Y=1, X \in B) = \int_B \eta(x) \mathrm{d}P_X(x).\end{aligned}$$

We deduce that $P_X[\{x \in \mathbb{R}^p : \eta(x) \neq \eta^{A^*}(A^*x)\}] = 0$; in particular, assumption 3 holds, as required.

A.5. Proof of proposition 2

We have

$$\begin{aligned}R(C^{A^*-\text{Bayes}}) &= \int_{\mathbb{R}^p \times \{0,1\}} \mathbb{1}_{\{C^{A^*-\text{Bayes}}(A^*x) \neq y\}} \mathrm{d}P(x, y) \\ &= \int_{\mathbb{R}^p} \eta(x) \mathbb{1}_{\{\eta^{A^*}(A^*x) < 1/2\}} \mathrm{d}P_X(x) + \int_{\mathbb{R}^p} \{1 - \eta(x)\} \mathbb{1}_{\{\eta^{A^*}(A^*x) \geqslant 1/2\}} \mathrm{d}P_X(x) \\ &= \int_{\mathbb{R}^p} \eta(x) \mathbb{1}_{\{\eta(x) < 1/2\}} \mathrm{d}P_X(x) + \int_{\mathbb{R}^p} \{1 - \eta(x)\} \mathbb{1}_{\{\eta(x) \geqslant 1/2\}} \mathrm{d}P_X(x) \\ &= R(C^{\text{Bayes}}),\end{aligned}$$

where we have used assumption 3 to obtain the penultimate equality.

A.6. Proof of proposition 3

Recall that $\sigma^2 := \theta(1-\theta)$. Let

$$F^*_{B_1}(s) = F^*_{B_1}(s,\theta) := \int_{-\infty}^{\infty} \exp(\mathrm{i}st)\,\mathrm{d}F_{B_1}(t) = \left[(1-\theta)\exp\left(-\frac{\mathrm{i}s\theta}{B_1^{1/2}\sigma}\right) + \theta\exp\left\{\frac{\mathrm{i}s(1-\theta)}{B_1^{1/2}\sigma}\right\}\right]^{B_1}.$$

Moreover, let $P(t) := \phi(t)p(t)/B_1^{1/2}$ and $Q(t) := \phi(t)q(t)/B_1^{1/2}$. By, for example, Gnedenko and Kolmogorov (1954), chapter 8, section 43, we have

$$\Phi^*(s) := \int_{\mathbb{R}} \exp(\mathrm{i}st)\,\mathrm{d}\Phi(t) = \exp\left(-\frac{s^2}{2}\right),$$

$$P^*(s) := \int_{\mathbb{R}} \exp(\mathrm{i}st)\,\mathrm{d}P(t) = -\frac{1-2\theta}{6B_1^{1/2}\sigma}\mathrm{i}s^3\exp\left(-\frac{s^2}{2}\right)$$

and

$$Q^*(s) := \int_{\mathbb{R}} \exp(\mathrm{i}st)\,\mathrm{d}Q(t) = -\frac{s}{2\pi B_1^{1/2}\sigma}\sum_{l\in\mathbb{Z}\setminus\{0\}} \frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l}\exp\left\{-\frac{1}{2}(s+2\pi B_1^{1/2}\sigma l)^2\right\}.$$

Thus

$$\begin{aligned} G^*_{B_1}(s) = G^*_{B_1}(s,\theta) &:= \int_{\mathbb{R}} \exp(\mathrm{i}st)\,\mathrm{d}G_{B_1}(t) = \Phi^*(s) + P^*(s) + Q^*(s) \\ &= \exp\left(-\frac{s^2}{2}\right) - \frac{1-2\theta}{6B_1^{1/2}\sigma}\mathrm{i}s^3\exp\left(-\frac{s^2}{2}\right) \\ &\quad - \frac{s}{2\pi B_1^{1/2}\sigma}\sum_{l\in\mathbb{Z}\setminus\{0\}} \frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l}\exp\left\{-\frac{1}{2}(s+2\pi B_1^{1/2}\sigma l)^2\right\}. \end{aligned}$$

Letting $c_2 > 0$ be the constant given in the statement of theorem 4 (in fact we assume without loss of generality that $c_2 > \pi$), we show that there is a constant $C' > 0$ such that, for all $B_1 \in \mathbb{N}$,

$$\sup_{\theta\in(0,1)} \sigma^3 \int_{-c_2 B_1^{1/2}\sigma}^{c_2 B_1^{1/2}\sigma} \left|\frac{F^*_{B_1}(s,\theta) - G^*_{B_1}(s,\theta)}{s}\right| \mathrm{d}s \leqslant \frac{C'}{B_1}. \tag{29}$$

To show inequality (29), write

$$\begin{aligned} \int_{-c_2 B_1^{1/2}\sigma}^{c_2 B_1^{1/2}\sigma} \left|\frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s}\right| \mathrm{d}s = \int_{-S_1}^{S_1} \left|\frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s}\right| \mathrm{d}s + \int_{S_1\leqslant|s|\leqslant S_2} \left|\frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s}\right| \mathrm{d}s \\ + \int_{S_2\leqslant|s|\leqslant c_2 B_1^{1/2}\sigma} \left|\frac{F^*_{B_1}(s) - G^*_{B_1}(s)}{s}\right| \mathrm{d}s, \end{aligned} \tag{30}$$

where $S_1 := B_1^{1/2}\sigma^{3/2}/\{32(3\theta^2 - 3\theta + 1)^{3/4}\}$ and $S_2 := \pi B_1^{1/2}\sigma$. Note that $S_1 \leqslant S_2/2$ for all $\theta \in (0,1)$.

We bound each term in equation (30) in turn. By Gnedenko and Kolmogorov (1954), theorem 1, section 4.1, there is a universal constant $C_3 > 0$, such that, for all $|s| \leqslant S_1$,

$$|F^*_{B_1}(s,\theta) - \Phi^*(s) - P^*(s)| \leqslant \frac{C_3}{B_1\sigma^3}(s^4 + s^6)\exp\left(-\frac{s^2}{4}\right).$$

Thus

$$\int_{-S_1}^{S_1} \left|\frac{F^*_{B_1}(s) - \Phi^*(s) - P^*(s)}{s}\right| \mathrm{d}s \leqslant \frac{C_3}{B_1\sigma^3}\int_{-\infty}^{\infty} (|s|^3 + |s|^5)\exp\left(-\frac{s^2}{4}\right)\mathrm{d}s = \frac{144C_3}{B_1\sigma^3}. \tag{31}$$

Moreover, observe that $(s+2\pi B_1^{1/2}\sigma l)^2 \geqslant s^2+2\pi^2 B_1\sigma^2 l^2$ for all $|s|\leqslant S_1$. Thus, for $|s|\leqslant S_1$,

$$\begin{aligned}\left|\frac{Q^*(s)}{s}\right| &\leqslant \frac{1}{2\pi B_1^{1/2}\sigma}\left|\sum_{l\in\mathbb{Z}\setminus\{0\}}\frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l}\exp\left\{-\frac{1}{2}(s+2\pi B_1^{1/2}\sigma l)^2\right\}\right| \\ &\leqslant \frac{\phi(s)}{\surd(2\pi)B_1^{1/2}\sigma}\int_{-\infty}^{\infty}\exp(-\pi^2 B_1\sigma^2 u^2)\,\mathrm{d}u = \frac{\phi(s)}{\surd 2\pi B_1\sigma^2}.\end{aligned}$$

It follows that

$$\int_{-S_1}^{S_1}\left|\frac{Q^*(s)}{s}\right|\mathrm{d}s \leqslant \frac{1}{\surd 2\pi B_1\sigma^2}. \qquad (32)$$

For $|s|\in[S_1,S_2]$, observe that

$$|F_{B_1}^*(s)| = \left[1-2\sigma^2\left\{1-\cos\left(\frac{s}{B_1^{1/2}\sigma}\right)\right\}\right]^{B_1/2} \leqslant \exp\left(-\frac{s^2}{8}\right).$$

Thus

$$\int_{S_1\leqslant|s|\leqslant S_2}\left|\frac{F_{B_1}^*(s)}{s}\right|\mathrm{d}s \leqslant \frac{2}{S_1^2}\int_{S_1}^{S_2}s\exp\left(-\frac{s^2}{8}\right)\mathrm{d}s \leqslant \frac{2^{13}}{B_1\sigma^3}. \qquad (33)$$

Now,

$$\int_{S_1\leqslant|s|\leqslant S_2}\left|\frac{\Phi^*(s)}{s}\right|\mathrm{d}s \leqslant \frac{2}{S_1^2}\int_0^{\infty}s\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s \leqslant \frac{2^{11}}{B_1\sigma^3}, \qquad (34)$$

and

$$\int_{S_1\leqslant|s|\leqslant S_2}\left|\frac{P^*(s)}{s}\right|\mathrm{d}s \leqslant \frac{1}{3S_1B_1^{1/2}\sigma}\int_0^{\infty}s^3\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s \leqslant \frac{2^6}{3\surd 2B_1\sigma^3}. \qquad (35)$$

To bound the final term, observe that, for all $|s|\in[S_1,S_2]$, since $(a+b)^2\geqslant(a^2+b^2)/5$ for all $|a|\leqslant|b|/2$, we have

$$\int_{S_1\leqslant|s|\leqslant S_2}\left|\frac{Q^*(s)}{s}\right|\mathrm{d}s \leqslant \frac{1}{2\pi B_1^{1/2}\sigma}\int_{S_1\leqslant|s|\leqslant S_2}\exp\left(-\frac{s^2}{10}\right)\int_{-\infty}^{\infty}\exp\left(\frac{-2\pi^2B_1\sigma^2u^2}{5}\right)\mathrm{d}u\,\mathrm{d}s \leqslant \frac{5}{4\pi B_1\sigma^3}. \qquad (36)$$

Finally, for $|s|\in[S_2,c_2B_1^{1/2}\sigma]$, note that

$$\begin{aligned}\int_{S_2\leqslant|s|\leqslant c_2B_1^{1/2}\sigma}\left|\frac{\Phi^*(s)+P^*(s)}{s}\right|\mathrm{d}s &\leqslant \frac{2}{S_2^2}\int_0^{\infty}s\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s + \frac{1}{3S_2B_1^{1/2}\sigma}\int_0^{\infty}s^3\exp\left(-\frac{s^2}{2}\right)\mathrm{d}s \\ &\leqslant \frac{1}{\pi^2B_1\sigma^3}\left(1+\frac{\pi}{3}\right). \qquad (37)\end{aligned}$$

To bound the remaining terms, by substituting $s=B_1^{1/2}\sigma u$, we see that

$$\int_{S_2}^{c_2B_1^{1/2}\sigma}\left|\frac{F_{B_1}^*(s)-Q_{B_1}^*(s)}{s}\right|\mathrm{d}s = \int_{\pi}^{c_2}\left|\frac{F_{B_1}^*(B_1^{1/2}\sigma u)-Q_{B_1}^*(B_1^{1/2}\sigma u)}{u}\right|\mathrm{d}u$$

$$
\begin{aligned}
&= \sum_{j=1}^{J} \int_{\pi(2j-1)}^{\pi(2j+1)} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| \mathrm{d}u \\
&\quad + \int_{\pi(2J+1)}^{c_2} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| \mathrm{d}u, \qquad (38)
\end{aligned}
$$

where $J := \lfloor (c_2 - \pi)/(2\pi) \rfloor$. Let

$$
\begin{aligned}
I_j &:= \int_{\pi(2j-1)}^{\pi(2j+1)} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma u) - Q_{B_1}^*(B_1^{1/2}\sigma u)}{u} \right| \mathrm{d}u \\
&= \int_{-\pi}^{\pi} \left| \frac{F_{B_1}^*\{B_1^{1/2}\sigma(v+2\pi j)\} - Q_{B_1}^*\{B_1^{1/2}\sigma(v+2\pi j)\}}{v+2\pi j} \right| \mathrm{d}v. \qquad (39)
\end{aligned}
$$

Observe that

$$
\begin{aligned}
F_{B_1}^*\{B_1^{1/2}\sigma(v+2\pi j)\} &= [(1-\theta)\exp\{-\mathrm{i}(v+2\pi j)\theta\} + \theta\exp\{\mathrm{i}(v+2\pi j)(1-\theta)\}]^{B_1} \\
&= \exp(-2\mathrm{i}\pi B_1 j\theta)[(1-\theta)\exp(-\mathrm{i}v\theta) + \theta\exp\{\mathrm{i}v(1-\theta)\}]^{B_1} \\
&= \exp(-2\mathrm{i}\pi B_1 j\theta) F_{B_1}^*(B_1^{1/2}\sigma v).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
Q_{B_1}^*\{B_1^{1/2}\sigma(v+2\pi j)\} &= -\frac{(v+2\pi j)}{2\pi} \sum_{l\in\mathbb{Z}\setminus\{0\}} \frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l} \exp\left\{-\frac{B_1\sigma^2}{2}(v+2\pi j+2\pi l)^2\right\} \\
&= \frac{(v+2\pi j)\exp(-2\mathrm{i}\pi B_1 j\theta)}{2\pi j} \exp\left(-\frac{B_1\sigma^2 v^2}{2}\right) \\
&\quad - \frac{(v+2\pi j)}{2\pi} \sum_{l\in\mathbb{Z}\setminus\{0,-j\}} \frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l} \exp\left\{-\frac{B_1\sigma^2}{2}(v+2\pi j+2\pi l)^2\right\}.
\end{aligned}
$$

But, for $v \in [-\pi, \pi]$,

$$
\begin{aligned}
\left| \frac{1}{2\pi} \sum_{l\in\mathbb{Z}\setminus\{0,-j\}} \frac{\exp(2\mathrm{i}\pi B_1 l\theta)}{l} \exp\left\{-\frac{B_1\sigma^2}{2}(v+2\pi j+2\pi l)^2\right\} \right| &\leqslant \frac{1}{2\pi} \sum_{m\in\mathbb{Z}\setminus\{0\}} \exp\left\{-\frac{B_1\sigma^2}{2}(v+2\pi m)^2\right\} \\
&\leqslant \frac{\exp(-B_1\sigma^2 v^2/10)}{2\pi} \sum_{m\in\mathbb{Z}\setminus\{0\}} \exp(-2\pi^2 B_1\sigma^2 m^2/5) \\
&\leqslant \frac{\exp(-B_1\sigma^2 v^2/10)}{\pi\{\exp(2\pi^2 B_1\sigma^2/5) - 1\}} \leqslant \frac{5\exp(-B_1\sigma^2 v^2/10)}{2\pi^3 B_1\sigma^2}.
\end{aligned}
$$

It follows that

$$
I_j \leqslant \int_{-\pi}^{\pi} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma v) - \{v/(2\pi j)+1\}\exp(-B_1\sigma^2 v^2/2)}{v+2\pi j} \right| \mathrm{d}v + \frac{5\surd 5}{\surd 2\pi^{5/2} B_1^{3/2}\sigma^3}. \qquad (40)
$$

Now

$$
\begin{aligned}
\int_{-\pi}^{\pi} \left| \frac{F_{B_1}^*(B_1^{1/2}\sigma v) - \exp(-B_1\sigma^2 v^2/2)}{v+2\pi j} \right| \mathrm{d}v &\leqslant \frac{1}{\pi j B_1^{1/2}\sigma} \int_{-\pi B_1^{1/2}\sigma}^{\pi B_1^{1/2}\sigma} \left| F_{B_1}^*(u) - \exp\left(-\frac{u^2}{2}\right) \right| \mathrm{d}u \\
&= \frac{1}{\pi j B_1^{1/2}\sigma} \int_{-S_3}^{S_3} \left| F_{B_1}^*(u) - \exp\left(-\frac{u^2}{2}\right) \right| \mathrm{d}u \\
&\quad + \frac{1}{\pi j B_1^{1/2}\sigma} \int_{S_3 \leqslant |u| \leqslant \pi B_1^{1/2}\sigma} \left| F_{B_1}^*(u) - \exp\left(-\frac{u^2}{2}\right) \right| \mathrm{d}u, \qquad (41)
\end{aligned}
$$

where $S_3 := B_1^{1/2}\sigma/\{5(2\theta^2-2\theta+1)\} \geqslant S_1$. By Gnedenko and Kolmogorov (1954), theorem 2, section 40, we have that

$$\frac{1}{\pi j B_1^{1/2}\sigma}\int_{-S_3}^{S_3}\left|F_{B_1}^*(u)-\exp\left(-\frac{u^2}{2}\right)\right|\mathrm{d}u \leqslant \frac{7}{6\pi j B_1\sigma^2}\int_{-S_3}^{S_3}|u|^3\exp\left(-\frac{u^2}{4}\right)\mathrm{d}u \leqslant \frac{56}{3\pi j B_1\sigma^2}. \qquad (42)$$

Moreover,

$$\frac{1}{\pi j B_1^{1/2}\sigma}\int_{S_3\leqslant|u|\leqslant\pi B_1^{1/2}\sigma}\left|F_{B_1}^*(u)-\exp\left(-\frac{u^2}{2}\right)\right|\mathrm{d}u \leqslant \frac{2}{\pi j S_3 B_1^{1/2}\sigma}\int_0^\infty u\left\{\exp\left(-\frac{u^2}{8}\right)+\exp\left(-\frac{u^2}{2}\right)\right\}\mathrm{d}u \leqslant \frac{50}{\pi j B_1\sigma^2}. \qquad (43)$$

Finally,

$$\frac{1}{2\pi j}\int_{-\pi}^{\pi}\frac{|v|}{|v|+2\pi j}\exp\left(-\frac{B_1\sigma^2 v^2}{2}\right)\mathrm{d}v \leqslant \frac{1}{2\pi^2 j^2}\int_0^\pi v\exp\left(-\frac{B_1\sigma^2 v^2}{2}\right)\mathrm{d}v \leqslant \frac{1}{2\pi^2 j^2 B_1\sigma^2}. \qquad (44)$$

By expressions (38)–(44), it follows that

$$\begin{aligned}\int_{S_2\leqslant|s|\leqslant c_2 B_1^{1/2}\sigma}\left|\frac{F_{B_1}^*(s)-Q_{B_1}^*(s)}{s}\right|\mathrm{d}s &\leqslant \frac{10\surd 5(J+1)}{\surd 2\pi^{5/2}B_1^{3/2}\sigma^3}+\frac{140}{\pi B_1\sigma^2}\sum_{j=1}^{J+1}\frac{1}{j}\\ &\leqslant \frac{10\surd 5(J+1)}{\surd 2\pi^{5/2}B_1^{3/2}\sigma^3}+\frac{140}{\pi B_1\sigma^2}\{1+\log(J+1)\}. \qquad (45)\end{aligned}$$

By expressions (30)–(37) and (45), we conclude that inequality (29) holds. The result now follows from theorem 4, by taking $c_1 = 1/(B_1^{1/2}\sigma)$, $C_1 = 1/(3B_1^{1/2}\sigma)$ and $S = c_2 B_1^{1/2}\sigma$ in that result.

References

Ailon, N. and Chazelle, B. (2006) Approximate nearest neighbours and the fast Johnson–Lindenstrauss transform. In *Proc. Symp. Theory of Computing*, pp. 557–563. New York: Association for Computing Machinery.

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

Blaser, R. and Fryzlewicz, P. (2015) Random rotation ensembles. *J. Mach. Learn. Res.*, **17**, 1–26.

Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984) *Classification and Regression Trees*. New York: Chapman and Hall.

Cannings, T. I. and Samworth, R. J. (2016) RPEnsemble: random projection ensemble classification. *R Package, Version 0.3*. Centre for Mathematical Sciences, University of Cambridge, Cambridge. (Available from `https://cran.r-project.org/web/packages/RPEnsemble/index.html`.)

Chikuse, Y. (2003) *Statistics on Special Manifolds*. New York: Springer.

Cook, R. D. (1998) *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Dasgupta, S. (1999) Learning mixtures of Gaussians. In *Proc. 40th A. Symp. Foundations of Computer Science*, pp. 634–644. Los Alamitos: Institute of Electrical and Electronics Engineers Computer Society.

Dasgupta, S. and Gupta, A. (2002) An elementary proof of the Johnson–Lindenstrauss Lemma. *Rand. Struct. Alg.*, **22**, 60–65.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Devroye, L. P. and Wagner, T. J. (1976) A distribution-free performance bound in error estimation. *IEEE Trans. Inform. Theory*, **22**, 586–587.

Devroye, L. P. and Wagner, T. J. (1979) Distribution-free inequalities for the deleted and hold-out error estimates. *IEEE Trans. Inform. Theory*, **25**, 202–207.

Durrant, R. J. and Kabán, A. (2013) Sharp generalization error bounds for randomly-projected classifiers. *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **28**, 693–701.

Durrant, R. J. and Kabán, A. (2015) Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Mach. Learn.*, **99**, 257–286.

Efron, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Ass.*, **70**, 892–898.

Esseen, C.-G. (1945) Fourier analysis of distribution functions: a mathematical study of the Laplace–Gaussian law. *Acta Math.*, **77**, 1–125.

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J., Feng, Y. and Tong, X. (2012) A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Statist. Soc.* B, **74**, 745–771.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Fix, E. and Hodges, J. L. (1951) Discriminatory analysis—nonparametric discrimination: consistency properties. *Technical Report 4*. US Air Force School of Aviation Medicine, Randolph Field.

Friedman, J. (1989) Regularized discriminant analysis. *J. Am. Statist. Ass.*, **84**, 165–175.

Gnedenko, B. V. and Kolmogorov, A. N. (1954) *Limit Distributions for Sums of Independent Random Variables*. Cambridge: Addison-Wesley.

Goeman, J., Meijer, R. and Chaturvedi, N. (2015) penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. *R Package Version 0.9-45*. Leiden University Medical Center, Leiden. (Available from `http://cran.r-project.org/web/packages/penalized/`.)

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W. and Lausen, B. (2015) ESKNN: ensemble of subset of K-nearest neighbours classifiers for classification and class membership probability estimation. *R Package Version 1.0*. Department of Mathematical Sciences, University of Essex, Colchester. (Available from `http://cran.r-project.org/web/packages/ESKNN/`.)

Gul, A., Perperoglou, A., Khan, Z., Mahmoud, O., Miftahuddin, M., Adler, W. and Lausen, B. (2016) Ensemble of a subset of *k*NN classifiers. *Adv. Data Anal. Classifcn*, 1–14.

Hall, P. and Kang, K.-H. (2005) Bandwidth choice for nonparametric classification. *Ann. Statist.*, **33**, 284–306.

Hall, P., Park, B. U. and Samworth, R. J. (2008) Choice of neighbour order in nearest-neighbour classification. *Ann. Statist.*, **36**, 2135–2152.

Hall, P. and Samworth, R. J. (2005) Properties of bagged nearest neighbour classifiers. *J. R. Statist. Soc.* B, **67**, 363–379.

Hastie, T., Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis. *Ann. Statist.*, **23**, 73–102.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hastie, T., Tibshirani, R., Narisimhan, B. and Chu, G.. (2015) pamr: Pam: prediction analysis for microarrays. *R Package Version 1.55*. Stanford University, Stanford. (Available from `http://cran.r-project.org/web/packages/pamr/`.)

Karatzoglou, A., Smola, A. and Hornik, K. (2015) kernlab: Kernel-based Machine Learning Lab. *R Package Version 0.9-20*. Telefonica Research, Barcelona. (Available from `http://cran.r-project.org/web/packages/kernlab/`.)

Khan, Z., Gul, A., Mahmoud, O., Miftahuddin, M., Perperoglou, A., Adler, W. and Lausen, B. (2015a) An ensemble of optimal trees for class membership probability estimation. In *Analysis of Large and Complex Data: Proc. Eur. Conf. Data Analysis, Bremen, July* (eds A. Wilhelm and H. A. Kestler). Berlin: Springer.

Khan, Z., Gul, A., Mahmoud, O., Miftahuddin, M., Perperoglou, A., Adler, W. and Lausen, B. (2015b) OTE: optimal trees ensembles for regression, classification and class membership probability estimation. *R Package Version 1.0*. (Available from `http://cran.r-project.org/web/packages/OTE/`.)

Larsen, K. G. and Nelson, J. (2016) The Johnson–Lindenstrauss lemma is optimal for linear dimensionality reduction. In *Proc. 43rd Int. Colloq. Automata, Languages and Programming*, pp. 1–11. Dagstuhl: Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Le, Q., Sarlos, T. and Smola, A. (2013) Fastfood—approximating kernel expansions in loglinear time. *J. Mach. Learn. Res. Wrkshp Conf. Proc.*, **28**, 244–252.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013) A general theory for nonlinear sufficient dimension reduction: formulation and estimation. *Ann. Statist.*, **41**, 221–249.

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, **86**, 316–342.

Liaw, A. and Wiener, M. (2014) randomForest: Breiman and Cutler's random forests for classification and regression. *R Package Version 4.6-10*. Merck, Kenilworth. (Available from `http://cran.r-project.org/web/packages/randomForest/`.)

Lopes, M. (2016) A sharp bound on the computation-accuracy tradeoff for majority voting ensembles. *Preprint arXiv:1303.0727v2*. University of California at Davis, Davis.

Lopes, M., Jacob, L. and Wainwright, M. J. (2011) A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*. Red Hook: Curran Associates.

Marzetta, T., Tucci, G. and Simon, S. (2011) A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Trans. Inform. Theory*, **57**, 6256–6271.

McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G. and Vanchinathan, H. P. (2014) LOCO: distributing ridge regression with random projections. *Preprint arXiv:1406.3469v2*.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C. and Lin, C.-C. (2015) e1071: Misc Functions of the Department of Statistics (e1071). *R Package Version 1.6-4*. Technical University of Vienna, Vienna. (Available from `http://cran.r-project.org/web/packages/e1071/`.)

Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, **40**, 2733–2763.

Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: another look at stability selection. *J. R. Statist. Soc.* B, **75**, 55–80.

Shin, S. J., Wu, Y., Zhang, H. H. and Liu, Y. (2014) Probability-enhanced sufficient dimension reduction for binary classification. *Biometrics*, **70**, 546–555.

Tibshirani, R., Hastie, T., Narisimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. In *Proc. Natn. Acad. Sci. USA*, **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narisimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, **18**, 104–117.

Trefethen, L. N. and Bau III, D. (1997) *Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (eds Y. C. Eldar and G. Kutyniok), pp. 210–268. Cambridge: Cambridge University Press.

Williams, C. K. I. and Barber, D. (1998) Bayesian classification with Gaussian processes. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1342–1351.

Witten, D. (2011) penalizedLDA: penalized classification using Fisher's linear discriminant. *R Package Version 1.0*. University of Washington, Seattle. (Available from `http://cran.r-project.org/web/packages/penalizedLDA/`.)

Witten, D. M. and Tibshirani, R. (2011) Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc.* B, **73**, 753–772.

Discussion on the paper by Cannings and Samworth

Christian Hennig (*University College London*) **and Cinzia Viroli** (*University of Bologna*)

Cannings and Samworth's random-projection (RP) ensemble is a fascinating new classification method for high dimensional data. One of us supervised a Master's project in the past in which ensemble classification was tried out based on plain random projections. This did not work very well. The authors of this paper provide us with some insight about this failure. Their approach to use 'competition winning' RPs addresses the issue successfully.

We believe that for full understanding of the performance of new methods it is required to find and understand some situations in which it does not work well. We believe that such situations exist for all classification methods.

We tried two approaches to find such situations. We present here all simulations that we ran; there is no selection bias.

Firstly, we were curious about how the RP classifiers would perform in the simulation set-ups in Hennig and Viroli (2016), which we used to test our quantile-based classifier (QC with versions QCG and QCS; see Hennig and Viroli (2016) for all details on the study). QC aggregates weighted distances to an optimally chosen quantile over all dimensions. The competitors were the centroid classfier CC (Tibshirani *et al.*, 2002), the median classifier MC (Hall *et al.*, 2009), linear discriminant analysis LDA, k-nearest neighbours $k$nn, the naive Bayes classifier n-Bayes, the support vector machine SVM, non-shrunken centroids NSC, penalized logistic regression stepPlr (Park and Hastie, 2008) and classification trees rpart (Breiman *et al.*, 1984).

Out of the study in Hennig and Viroli (2016) we used only two different combinations of n and p and two percentages of informative variables (IVs). The base set-ups were run with either dependent (set-ups 2 and 4) or independent (set-ups 1 and 3) variables. Class sizes were balanced except in set-up 4, where class proportions were 0.75 and 0.25.

QC, CC, MC, $k$nn and n-Bayes aggregate information over all variables. The other competitors including the RP classifier aim at lower dimensional subspaces in which classification works best. Such methods may be advantageous if classification information is concentrated on a lower dimensional subspace. Our set-ups with 100% IVs can be expected to favour the former methods, whereas with 10% IVs the latter approaches could do better. Cannings and Samworth's assumption 3 is never fulfilled here with d as small as 5.

Misclassification rates are shown in Tables 7 and 8. In set-up 1, RP with LDA or $k$nn does well. In set-up 3, all versions of RP are competitive for 100% IVs; they do somewhat worse for 10% IVs. In set-ups 2 and 4 with independent skewed distributed variables, all RP classifiers perform badly. Often they deliver the worst performances out of all methods (besides being slowest). For 100% IVs they were surprisingly

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Random projection ensemble classification: supplementary material'.

Table 7. Misclassification rates for set-ups 1 and 2 (with standard errors in parentheses)†

IV	*Rates for set-up 1, dependent ID symmetric variables*				*Rates for set-up 2, independent ID asymmetric variables*			
	n = 50, p = 100		*n = 500, p = 500*		*n = 50, p = 100*		*n = 500, p = 500*	
	100%	*10%*	*100%*	*10%*	*100%*	*10%*	*100%*	*10%*
QCG	11.2 (0.6)	41.6 (0.6)	1.2 (0.1)	19.0 (0.2)	25.7 (1.0)	44.3 (0.5)	0.0 (0.0)	6.3 (0.1)
QCS	11.1 (0.5)	41.2 (0.6)	1.2 (0.1)	18.5 (0.2)	21.3 (0.8)	41.5 (0.6)	0.0 (0.0)	6.3 (0.1)
CC	13.5 (0.9)	42.2 (0.6)	0.7 (0.0)	19.1 (0.3)	42.7 (0.5)	44.0 (0.5)	22.4 (0.2)	46.0 (0.2)
MC	10.3 (0.4)	40.9 (0.7)	1.1 (0.0)	18.1 (0.2)	34.3 (0.7)	44.8 (0.4)	5.2 (0.1)	38.5 (0.2)
LDA	23.6 (0.7)	42.9 (0.5)	1.2 (0.1)	19.5 (0.7)	44.0 (0.4)	44.9 (0.4)	46.8 (0.2)	48.3 (0.2)
*k*nn	13.8 (0.6)	44.0 (0.5)	1.1 (0.1)	32.5 (0.6)	45.6 (0.4)	45.3 (0.4)	45.6 (0.2)	48.5 (0.1)
n-Bayes	21.0 (0.7)	43.4 (0.5)	5.6 (0.2)	35.9 (0.8)	42.7 (0.5)	45.1 (0.4)	28.3 (0.2)	46.7 (0.2)
SVM	7.7 (0.5)	38.3 (0.7)	1.3 (0.1)	8.7 (0.1)	42.4 (0.5)	44.2 (0.4)	19.2 (0.2)	46.0 (0.2)
NSC	26.4 (0.7)	41.4 (0.6)	1.3 (1.1)	12.8 (0.2)	44.8 (0.5)	44.3 (0.4)	35.8 (0.3)	46.3 (0.2)
stepPlr	5.3 (0.3)	36.7 (0.8)	0.4 (0.0)	8.4 (0.1)	42.3 (0.5)	44.4 (0.4)	27.8 (0.2)	46.8 (0.2)
rpart	40.1 (0.6)	43.2 (0.5)	40.6 (0.2)	41.0 (0.2)	42.6 (0.5)	44.3 (0.4)	23.1 (0.3)	21.7 (0.3)
RP-LDA$_2$	8.3 (0.5)	41.0 (0.7)	0.7 (0.0)	15.4 (0.2)	44.2 (0.7)	50.2 (0.6)	24.4 (0.2)	45.9 (0.2)
RP-LDA$_5$	8.0 (0.5)	41.6 (0.7)	0.6 (0.0)	14.9 (0.2)	44.6 (0.8)	49.5 (0.7)	23.3 (0.2)	46.0 (0.2)
RP-QDA$_2$	11.4 (0.6)	45.0 (0.7)	1.5 (0.1)	26.5 (0.8)	46.1 (0.7)	50.1 (0.7)	24.4 (0.3)	46.9 (0.2)
RP-QDA$_5$	12.4 (0.7)	45.2 (0.6)	2.2 (0.2)	32.2 (1.0)	46.5 (0.7)	49.4 (0.7)	25.2 (0.4)	47.4 (0.2)
RP-*k*nn$_2$	8.9 (0.5)	42.4 (0.7)	0.8 (0.0)	24.5 (0.4)	45.7 (0.7)	50.2 (0.6)	36.7 (0.2)	49.2 (0.2)
RP-*k*nn$_5$	8.4 (0.5)	41.7 (0.7)	0.7 (0.0)	18.1 (0.3)	45.3 (0.6)	50.7 (0.7)	30.0 (0.2)	47.9 (0.2)

†ID, ‘identically distributed’; 100 replications, i.e. training and test sets.

more competitive than for 10% IVs. (Note that for $n = 500$ and $p = 500$ in set-up 4 they always classified all observations to the bigger class.)

One could argue that high dimensional situations in which all variables are informative and (near) independent may be very rare, so this may not be seen as a big problem for the RP classifier. The difficulty to deal with the set-ups with 10% IVs seems to be a more important issue.

Secondly, we wondered whether there could be a problem caused by running the base classifier on ‘competition winning’ projections disregarding potential dependence between these. In case there is some dominating information represented in many dependent variables, we suspected that other useful information for classification could be ignored by the RP classifiers. To test this, we ran a simulation (set-up 5) with $n = 2 \times 200$ (200 observations in each class) and $p = 200$ variables $X_1, \ldots, X_p$. For class $i, j = 1, \ldots, 10$, $X_j \sim \mathcal{N}(a_{ij}, 1)$ independently with $a_{1j} = 0$, $a_{21} = a_{22} = 1.5$ and $a_{2k} = 0.3$ for $k = 3, \ldots, 10$. With $X_j^* \sim \mathcal{N}(0, 1)$ for $j = 11, \ldots, 200$ we used $X_k = 0.9X_1 + 0.1X_k^*$ for $k = 11, \ldots, 100$, $X_k = 0.9X_2 + 0.1X_k^*$ for $k = 101, \ldots, 150$ and $X_k = X_k^*$ for $k > 150$. The idea was that the classification information in X_1 and X_2 should be present also in many other variables, ‘masking’ the independent information in $X_3, \ldots, X_{10}$. Results are shown in Table 9. The RP classifiers, to our surprise, performed excellently here. If this set-up is a problem for them, it proved to be even more of a problem for almost all competitors.

A further interesting issue would be to give guidelines for when, in a practical situation, an RP classifier (and which) should be used. Also we wonder about how situations could be recognized in which assumption 2 is fulfilled (or not).

Overall this is a fascinating paper and it is a pleasure for us to propose the vote of thanks.

David Hand (*Imperial College London*)

I should like to begin by congratulating Cannings and Samworth on a very impressive piece of work. It combines extensive and revealing empirical demonstrations with deep and powerful theory and has implications beyond the application that is described in the paper.

The idea of random-projection ensemble classifiers has been previously explored, for example, by Schclar and Rokach (2009). However, this present paper explores the mathematical foundations and empirical comparisons in considerably greater detail and rigour than Schclar and Rokach (2009) did.

Table 8. Misclassification rates for set-ups 3 and 4 (with standard errors in parentheses)†

IV	*Rates for set-up 3, dependent not ID variables*				*Rates for set-up 4, beta not ID variables with unbalanced classes*			
	$n=50$, $p=100$		*$n=500$, $p=500$*		*$n=50$, $p=100$*		*$n=500$, $p=500$*	
	100%	*10%*	*100%*	*10%*	*100%*	*10%*	*100%*	*10%*
QCG	21.0 (1.0)	43.9 (0.5)	0.2 (0.0)	24.9 (0.3)	6.5 (0.5)	23.4 (0.2)	0.0 (0.0)	5.4 (0.2)
QCS	24.2 (0.8)	43.9 (0.4)	0.4 (0.0)	28.6 (0.3)	5.6 (0.5)	22.9 (0.3)	0.0 (0.0)	4.6 (0.2)
CC	19.6 (0.5)	44.0 (0.5)	0.4 (0.0)	24.9 (0.2)	25.6 (0.4)	26.3 (0.4)	29.6 (0.3)	31.3 (0.3)
MC	20.3 (0.6)	42.8 (0.5)	0.8 (0.0)	28.5 (0.2)	19.3 (0.5)	25.5 (0.3)	1.3 (0.1)	26.2 (0.2)
LDA	31.5 (0.7)	43.5 (0.5)	41.7 (0.4)	47.4 (0.2)	25.1 (0.3)	25.6 (0.3)	46.6 (0.2)	46.9 (0.2)
*k*nn	30.7 (0.6)	44.8 (0.4)	5.4 (0.2)	43.1 (0.2)	26.1 (0.5)	26.7 (0.5)	18.8 (0.3)	25.2 (0.1)
n-Bayes	31.6 (0.8)	43.5 (0.5)	4.8 (0.1)	38.8 (0.2)	4.1 (0.3)	23.1 (0.2)	0.0 (0.0)	8.9 (0.3)
SVM	20.9 (0.7)	43.7 (0.4)	0.5 (0.0)	27.1 (0.3)	24.0 (0.0)	24.0 (0.0)	22.3 (0.2)	25.0 (0.0)
NSC	25.2 (0.6)	41.7 (0.5)	0.4 (0.0)	18.9 (0.2)	24.0 (0.0)	24.0 (0.0)	25.0 (0.0)	25.0 (0.0)
stepPlr	21.8 (0.7)	43.7 (0.5)	0.6 (0.0)	28.2 (0.3)	24.0 (0.0)	24.0 (0.0)	25.7 (0.1)	25.0 (0.0)
rpart	30.5 (0.9)	38.8 (0.8)	3.9 (0.1)	2.8 (0.1)	16.5 (0.6)	21.4 (0.7)	2.0 (0.1)	4.4 (0.2)
RP-LDA$_2$	21.5 (0.6)	45.2 (0.6)	0.7 (0.0)	26.7 (0.3)	34.8 (1.0)	36.8 (1.0)	25.0 (0.0)	25.0 (0.0)
RP-LDA$_5$	21.8 (0.6)	45.9 (0.6)	0.6 (0.0)	26.0 (0.3)	36.9 (1.0)	40.1 (1.0)	25.0 (0.0)	25.0 (0.0)
RP-QDA$_2$	21.6 (0.7)	46.4 (0.6)	0.8 (0.0)	27.5 (0.3)	30.1 (0.9)	34.3 (1.2)	25.0 (0.0)	25.0 (0.0)
RP-QDA$_5$	22.2 (0.7)	46.8 (0.7)	0.6 (0.0)	26.9 (0.3)	27.0 (0.8)	33.8 (0.8)	25.0 (0.0)	25.0 (0.0)
RP-*k*nn$_2$	22.6 (0.7)	46.9 (0.6)	1.0 (0.1)	38.8 (0.2)	25.3 (0.6)	26.3 (0.5)	25.0 (0.0)	25.0 (0.0)
RP-*k*nn$_5$	21.4 (0.7)	46.2 (0.7)	0.7 (0.0)	32.5 (0.3)	30.3 (0.8)	34.9 (1.0)	25.0 (0.0)	25.0 (0.0)

†ID, ‘identically distributed’; 100 replications, i.e. training and test sets.

My questions relate mainly to the performance of random-projection methods. In general, the assessment of a method tells us how good it is, whether it is ‘good enough’ for some application, whether it is better than alternatives, and in what circumstances it has these properties. However, as I have shown in the past (Hand, 2006), evaluations can often be misleading.

Duin (1996) has observed that

> ‘In comparing classifiers one should realize that some classifiers are valuable because they are heavily parameterized and thereby offer a trained analyst a large flexibility in integrating his problem knowledge in the classification procedure. Other classifiers on the contrary, are very valuable because they are entirely automatic and do not demand any user parameter adjustment. As a consequence they can be used by anybody’.

The point is that comparisons are not really of the methods *per se*, but rather of the combination of method, user and the particular problem. The point was also elegantly illustrated by Bruce Hoadley in his ‘ping-pong theorem’ (Hoadley, 2001). He wrote:

> ‘This theorem says that if we revealed to Professor Breiman the performance of our best model and gave him our data, then he could develop an algorithmic model using random forests, which would outperform our model. But if he revealed to us the performance of his model, then we could develop a segmented scorecard, which would outperform his model.’

The point is also manifest as the ‘not invented here’ effect, in which studies comparing newly invented classification methods typically show that the new method is superior to existing methods—at least partly because the new method is, by definition, being applied by people who are particularly expert in that method and may be less so in applying the competitor methods.

The authors comment that

> ‘the flexibility offered by the random-projection ensemble classifier... allows the practitioner to adapt the method to work well in a wide variety of problems’,

but one might interpret that as meaning that an expert in the method can choose a variant so that it does

Table 9. Misclassification rates for set-up 5 (with standard errors in parentheses; 100 replications)

IV	*Rates for set-up 5, n = 200, p = 200*
QCG	16.7 (0.2)
QCS	17.1 (0.2)
CC	15.1 (0.2)
MC	17.5 (0.2)
LDA	24.4 (0.3)
*k*nn	17.1 (0.2)
n-Bayes	15.3 (0.2)
SVM	14.2 (0.2)
NSC	15.1 (0.2)
stepPlr	16.2 (0.2)
rpart	23.4 (0.2)
RP-LDA$_2$	14.3 (0.2)
RP-LDA$_5$	14.2 (0.2)
RP-QDA$_2$	14.1 (0.2)
RP-QDA$_5$	14.5 (0.2)
RP-*k*nn$_2$	14.4 (0.2)
RP-*k*nn$_5$	14.5 (0.2)

well in comparative evaluations. Perhaps it would have been fairer also to use several versions of the other classifiers in the comparisons. For example, the *k*nn classifiers could include weighting training points according to distance from the test point, as mentioned in the paper, and optimal choice of metric to define 'nearest', and there are now (after about 80 years of use and development) many variants of linear discriminant analysis.

More generally, although I appreciate that the vast number of classification methods which now exist means that comparative studies must make choices, that does inevitably open one to criticism. I would have liked to know how the method compared with other simpler methods as well as more elaborate methods. For example, although the basic naive Bayes classifier (assuming feature independence) is typically not the best in classification performance comparisons, it is usually among the best. Similarly, Holte (1993) showed that a method based on the single best predictor also typically did reasonably well. And, at the other extreme, methods which seem to be winning classification competitions at the moment seem to use variants of gradient boosting. State of the art comparisons should include these.

In a complementary vein, I am obliged to ask how you chose the particular data sets that you used in your comparisons. Generalizing from any such choices is always risky, and it could be that other choices would have given results less complimentary to random-projection methods. In my view, the key issue here is not that a particular method did well on a particular data set, but rather what properties of the data (and problem) led to the method's doing well. I accept that it is not easy to draw such conclusions (see Jamain (2004) and Jamain and Hand (2008)) but knowing this would be more useful than results on *ad hoc* data set choices and comparisons. Having said all that, I was pleased that the authors noted that, although coming top may be desirable, it is almost as attractive that a classifier is always quite good—a form of robustness.

The paper is very solidly based on the misclassification rate as the measure of performance. Is it really appropriate to assume that the two types of misclassification error are equally serious? More generally, it might be that entirely different but perhaps more relevant criteria give quite different results. This was illustrated, using the ionosphere data as it happens, by Thomas Benton (Benton (2002), pages 155–156), who produced a scatter plot based on two of the ionosphere data set variables, standardized to unit variance, indicating the directions which maximized the gradient of the area under the receiver operating characteristic curve, the proportion correctly classified, the log-likelihood and the standardized difference between the class means. The gradients pointed in different directions, with those for the area under the receiver operating characteristic curve and proportion correctly classified lying at an angle of almost 60° to each other. Classifiers built by using these two measures would yield very different results.

It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

Wenyang Zhang (*University of York*)
I congratulate Professor Samworth and Dr Cannings for such a brilliant paper. I believe that it will have much influence on high dimensional classification and will stimulate many further researches in this topic.

Classification is an important topic in data analysis. As the authors rightly point out, direct use of the traditional approaches, such as linear discriminant analysis or the k-nearest-neighbour classifier, may not work when the dimension of the data is high. In this paper, to deal with high dimensional classification, the authors project the high dimensional data to a lower dimensional space in a very clever way and construct an interesting classifier based on random projections sampled from the set of all projections.

For high dimensional classification, a natural approach would be based on some kind of penalized generalized linear type models, such as penalized logistic regression. The results presented in Table 1 for model 1 in the simulation study section for the penalized logistic regression seem a little strange. I understand that these results were obtained by using an R package; however, it seems to me that the tuning parameters used may not be optimal. Is there any explanation why the proposed classifier is better than the penalized logistic regression based classifier?

The projection of high dimensional data to a lower dimension space is somewhat equivalent to some features of the high dimensional data, and the distribution, based on which the random projections are sampled from the set of all projections, in the construction of the proposed classifier would act like a feature selection. If this is so, would the classifier proposed benefit significantly from a very careful selection of the distribution? Alternatively, would a weighted average be better than the simple average in expression (1)? Of course, the weights would depend on both n and p.

On the theoretical front, what kind of role does the p play in the classifier proposed? Would the classifier work for ultrahigh dimensional cases, say for example $\log(p) = O(n^{\alpha})$? Is d fixed? If it is, I guess that assumption 3 would be difficult to satisfy for ultrahigh dimensional cases.

For ultrahigh dimensional cases, the ideas developed in Li *et al.* (2015) and Ke *et al.* (2016) can also be used for classification. Their approach is based on penalized generalized linear type models though.

L. Anderlucci, A. Montanari and F. Fortunato (*University of Bologna, Italy*)
The paper is very motivating: the introduction of random projections (RPs) in the context of ensemble classifiers enables us to improve classification accuracy while extending to the high dimensional context methods originally developed for low dimensional data. However, a still open issue is the understanding of the properties of the variable ranking induced by the RP ensemple classifier. Although such a classifier highly improves the classification accuracy, it does not enable us to identify the variables with the highest discriminative power, like a single classifier does.

Inspired by the random-forest process for feature selection, our idea is to adjust the ensemble based on RP classifiers to keep the information on variable importance.

The idea is to detect the variables that mostly contribute to the best RP solution within each of the B_1 blocks of projections. Specifically, the input features are ranked according to their relative importance, measured through a specific coefficient, called the *variable importance in projection* (VIP).

Following Montanari and Lizzani (2001), for the ith variable the *importance coefficient* CI is defined as

$$\mathrm{CI}_{ib_1} = \sum_{j=1}^{d} \frac{|a_{ijb_1}| s_i}{\sqrt{\sum_{l=1}^{p} (a_{ijb_1} s_i)^2}} \qquad b_1 = 1, \ldots, B_1$$

where a_{ijb_1} indicates the attribute i coefficient in the jth vector of the d-dimensional RP solution and s_i the variability of each attribute. The VIP for feature i is then obtained as

$$\mathrm{VIP}_i = \operatorname*{median}_{b_1=1,\ldots,B_1} \mathrm{CI}_{ib_1}.$$

The $p-h$ variables that present the smallest values for the VIP coefficient are deemed not to contribute to the definition of the RP ensemble solution and, thus, can be removed. Our proposal explores all the possible solutions and retains the first h variables that minimize the test error estimate.

The VIP criterion has been tested in a Monte Carlo simulation study and in real data applications, focusing on the ability both to recover the actually important features and to perform an accurate classification. Results showed that adjusting the RP ensemble classifier with the VIP information for feature selection preserves the classification accuracy. In addition, the understanding of the classification problem is enhanced by providing a ranking of the features in terms of their discriminative power.

F. Fortunato (*University of Bologna*)
It is a great pleasure to comment on such a thought-provoking paper that motivated me to investigate further the discriminant analysis problem using random-projection (RP) ensembles.

Whereas Cannings and Samworth's idea of ensemble classification is deeply rooted in machine learning, the combination of such a strategy with RPs is definitely original.

My research in this context has aimed at investigating the ensemble post-pruning issue, as several studies have shown that having a large number of models in an ensemble could produce *redundancy*.

Specifically, I focused my attention on ways to decrease the number of classifiers in the ensemble while enhancing accuracy. Studying the characteristics of a *good* subset of classifiers, I noted that the binomial distribution $B_i(n,\pi)$ is not appropriate to describe the ensemble accuracy. In fact, in spite of the independence of the RPs, the assumption of independent classifiers is not realistic as they have been trained on the very same data. To account for the intraclassifiers association, I proposed to use a natural generalization of the binomial distribution to dependent binary data: the multiplicative binomial (MB) distribution, introduced by Altham (1978),

$$P(X=x)=\frac{\binom{n}{x}\psi^x(1-\psi)^{n-x}\omega^{(n-x)x}}{\sum_{i=0}^{n}\binom{n}{i}\psi^i(1-\psi)^{n-i}\omega^{(n-i)i}}.$$

Here, $0<\psi<1$ is the marginal probability of success and $\omega>0$ is the new parameter which governs the dependence between the binary responses: $\omega<1$ describes positively related responses, whereas $\omega>1$ a negative global relationship. Results coming from a broad simulation study confirm that the MB should be preferred to both the binomial and the beta–binomial models. The MB model, in fact, always seems to characterize and predict the classification accuracy of an ensemble of classifiers better.

Combining the idea of using the MB as the reference model for ensemble accuracy and my result on such a distribution,

$$\omega>1\Rightarrow\psi>\pi,$$

a simple ensemble selection algorithm (ESA) has been devised. This technique, starting from a single classifier ensemble E, at each step adds to the existing ensemble the ith classifier that is most similar to E in terms of accuracy and, at the same time, that provides the highest gain in terms of ω. Results of applying two different pruning models (the ESA and a multiobjective genetic algorithm), as well as the RP ensemble classifer on the same RP ensemble, demonstrate that the three are comparable in terms of accuracy rates and the ESA, without loss of accuracy (and actually sometimes doing even better) always tends to use a very small number of individual classifiers.

Frank Critchley (*The Open University, Milton Keynes*)
This is a wonderful paper, elegantly written and cogently argued, whose ultimate influence will I am sure range well beyond its current domain. I have some questions and comments, which I hope may suggest some fruitful ways forward.

(a) *Assumptions*: how far, and in which directions, might assumptions helpfully be relaxed (especially, assumption 3, which appears not to be necessary in one of the examples)?
(b) *Multiple classifiers*: might results usefully be combined between, as well as within, classifiers (especially, (when) is there useful information in disagreements between them)?
(c) *The Johnson–Lindenstrauss lemma*: might this existence result be used any more directly constructively (especially, its proof)?
(d) *Parameter choice*: avoiding 'one size fits all' guidance, might key method parameters ($\mathcal{D}$, (B_1, B_2), {base classifiers}, . . .) helpfully be chosen in terms of (n, p) and, both scientific and computational, context?
(e) *Marginal standardization of the p variables*: for any base classifier, this would, of itself, ensure

invariance to separate affine changes in their, generally disparate, scales of measurement. It could also help to reduce any undue influence of, say, variables with relatively large variances. Performed robustly, it would also highlight marginal outliers. That said, robustness may not be such a big issue here, as we discuss next.

(f) *Robustness*: correct classification of a data point varying naturally with projection, it may be hoped that the random-projection (RP) approach will prove relatively robust to outlying or 'rogue' observations x, and, to incorrect recording of class membership, y, in the training data.

(g) *Unsupervised classification*: might at least the spirit of RP be extended to the unsupervised case? Replacing estimated test error by an appropriate index of 'cluster structure' of an automatically identified subspace, invariant co-ordinate selection provides one possibility here, given the remarkable fact that this exploratory methodology can recover Fisher's linear discriminant subspace without knowing group membership.

(h) *Potential synergies*: what might RP and other methodologies have to offer each other? Sufficient dimension reduction is one natural candidate 'partner' here. My colleagues, Radka Sabolová and Paul Marriott, will comment further on this in their written contribution.

John T. Kent (*University of Leeds*)

Classic discriminant methods such as Fisher's linear discriminant rule are equivariant under affine transformations but require that n is not too large relative to p. Regularized methods may work well in high dimensions, especially $p > n$, but can only be equivariant under a smaller group, e.g. orthogonal transformations. Tonight's elegant randomized projection (RP) method is orthogonally equivariant (provided that the base classifier is) since it relies on the simulation of uniformly distributed directions on a unit sphere.

To test the limits of the RP method we can look for situations with low discriminatory power. Consider a toy Gaussian example in $p=2$ dimensions, where the two groups have common covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and let the vector δ denote the difference between the means.

The discriminatory power of Fisher's rule can be characterized by the squared Mahalanobis distance $D^2 = \delta^{\mathrm{T}} \Sigma^{-1} \delta$. After projecting onto a one-dimensional vector a, the discriminatory power drops to $D_a^2 = (a^{\mathrm{T}}\delta)^2/(a^{\mathrm{T}}\Sigma a)$, which ranges between 0, if $a \perp \delta$, and D^2, if $a \propto \Sigma^{-1}\delta$. Let $\theta \in [0, \pi]$ denote the angle between the random projection a and δ, and consider two cases for p.

(a) Case 1: $\rho = 0$. Then $D_a^2/D^2 = \cos^2(\theta)$.

(b) Case 2: $\rho = 0.99$ and δ lies in the direction of the smaller principal axis. Then

$$D_a^2/D^2 = \frac{\cos^2(\theta)}{\cos^2(\theta) + 199 \sin^2(\theta)}.$$

RPs are much less likely to see the signal in case 2; see Fig. 4. In particular, $D_a^2/D^2 \geqslant 0.5$ for 50% of random angles θ in case 1, but for only 5% of angles in case 2.

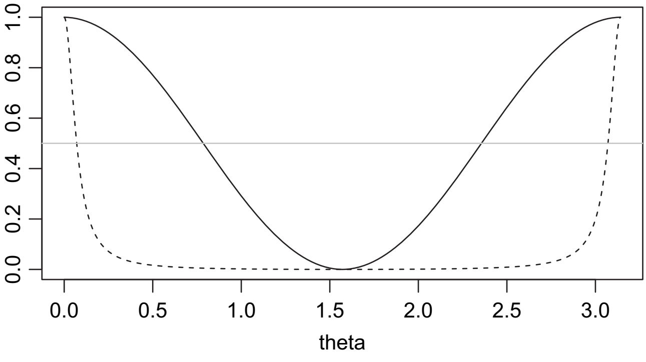


Fig. 4. Relative discriminatory power, as a function of the angle $\theta \in [0, \pi]$ for case 1 (———) and case 2 (------)

I am not sure whether to be concerned that RP can be broken in certain circumstances or reassured that such an extreme value of ρ is needed. I suspect that extreme correlations are prohibited in the small print of the regularity conditions. Perhaps a more nuanced reaction is to note that the simulated examples in the paper involve only moderate correlation and that it is not immediately obvious how the vector δ is oriented relative to the smallest principal axis. Further, it is not clear how much worse such effects will be in higher dimensions.

Yining Chen (*London School of Economics and Political Science*) **and Rajen D. Shah** (*University of Cambridge*)

We congratulate the authors for this interesting paper which introduces an important ensemble method for random projections in classification problems. We shall limit our comments to the procedure of selecting random projections and aggregating the results.

The basic procedure, as stated in Section 3, involves forming $B := B_1 \times B_2$ random projections of the data. A base classifier (e.g. k-nearest neighbours) is trained on each of these B projected versions of the data. The resulting classifiers are then grouped consecutively into blocks of size B_2, where we pick and then average those with the lowest training or leave-one-out (LOO) cross-validation error from each group and discard the rest. However, the blocking strategy perhaps does not make full use of the information from the training or LOO estimates whose construction is usually the most computationally intensive part of the procedure. Indeed, grouping base classifiers consecutively is somewhat arbitrary: the distribution of the ensemble classifier, conditional on the data and the set of random projections, is unchanged when permuting the list of classifiers. Therefore, one can construct new ensemble classifiers resulting from multiple random groupings with little extra computational cost. Here each new classifier is still based on the B base classifiers, but we instead randomly permute the order of the base classifiers before grouping them into blocks consecutively. By aggregating these new classifiers by a simple majority vote, we form a final classifier, which could potentially remove some of the variance resulting from the randomness of the grouping.

To examine the performance, we applied both the original method and the variant to four real data sets by using k-nearest neighbours with different training set sizes and setting $B = 1000$ and $B_2 = 50$. Results are reported in Fig. 5. As expected, the proposed variant with multiple random grouping gives slightly improved performance.

More generally, we could think of the training or LOO predictions from the base classifiers as covariates of the new training data for a further classifier; an approach known as stacking or blending (Wolpert, 1992; Breiman, 1996). We looked at forming a final classifier via regression of the class labels on the LOO predictions of k-nearest neighbours using l_1-penalized logistic regression with a non-negativity constraint on the coefficients. This can be viewed as a data-driven way of forming a weighted average of B classifiers

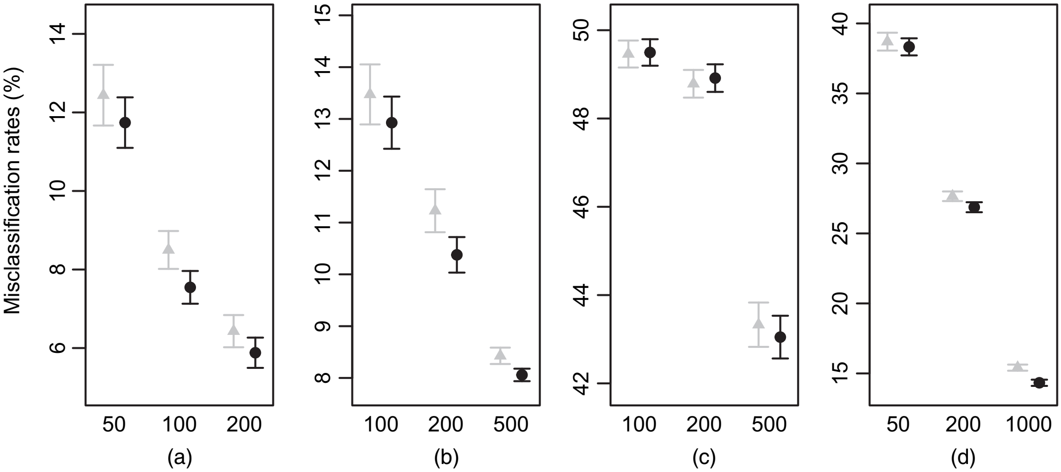


Fig. 5. Misclassification rates and the corresponding confidence intervals of the original random-projection ensemble classifier (▲) and the multiple random-grouping approach (●) on four real data sets (considered by the authors in Section 6.2) with various training set sizes and $(B, B_2) = (1000, 50)$: (a) ionosphere; (b) musk; (c) hill–valley; (d) eye state

Table 10. Estimated misclassification rates and the corresponding standard errors of various classifiers for the eye state data

Classifier	*Misclassification rate (%)*
k-nn	$14.45_{0.16}$
RP-$k$nn$_5$, $B=25000$	$13.54_{0.19}$
RP-$k$nn$_5$ with stacking, $B=500$	$12.86_{0.08}$
RP-$k$nn$_5$ with stacking, $B=5000$	$11.35_{0.07}$

on the projected versions of the data. Results on the eye state data (see Section 6.2.1, where RP-$k$nn$_5$ performed the best) with $n=1000$ are shown in Table 10. This suggests that some slightly more data-driven variants of the aggregation procedure that were used in the paper may lead to further improved performance in some settings, even with a smaller B.

The following contributions were received in writing after the meeting.

Amir Ahmad (*United Arab Emirates University, Al Ain*)
I congratulate Cannings and Samworth for this interesting paper. The paper discusses the use of random projections for classifier ensembles for high dimensional classification. Microarray data sets are examples of high dimensional data sets. It would be interesting if the authors could show some results on these data sets.

The authors propose selecting some random projections from the pool of random projections for the final results. Readers will benefit if the authors can show the effect of this step on real data sets by experiments.

Wicher Bergsma and Haziq Jamil (*London School of Economics and Political Science*)
Cannings and Samworth present impressive theoretical results. However, we are not yet convinced about their practical use: as summarized in Table 11, we ran our own favourite classifier—Gaussian process regression using fractional Brownian motion, GPR-FBM—and for five of the eight data sets we obtained results better than random-projection (RP) ensembles. Furthermore, our preliminary analyses indicate that RP ensembles worsen GPR-FBM classification, but this could be due to the small B_1 and B_2 we chose because of time constraints ($B_1=30$ and $B_2=5$). Thus, although RP ensemble methods can demonstrably improve frequently poor methods such as linear discriminant analysis, LDA, and $k$nn, we wonder whether they can improve good methods. If not, what then is the advantage of using RP ensembles?

It appears to us that there may be a mismatch between theory and practice. Theory tells us that the curse of dimensionality is a problem for high dimensional regression and classification; for example, according to Hastie and Tibshirani (1986), page 305, 'the chief motivation for the additive model' is that 'it is well known that smoothers break down in higher dimensions [because] the curse of dimensionality takes its toll'. However, in view of the success of, for example, support vector machines and GPR, and the results in Table 11, it seems to us that in practice smoothers do *not* break down in high dimensions. So it might be wondered, is the curse of dimensionality a straw man undeserving of the broad attention it is receiving?

Our own GPR methodology and associated R package will be made available soon via arXiv. In particular, we shall propose a flexible empirical Bayes methodology based on the Fisher information for the regression function, which in a well-defined sense can improve on Tikhonov regularization, and can further improve some of the GPR results in Table 11.

Finally, our results were obtained by fitting the model

$$y_i=f(x_i)+\varepsilon_i, \qquad y_i\in\{0,1\}, \quad x_i\in\mathbb{R}^p, \quad f\sim\mathrm{GP}(0,K), \quad K:\mathbb{R}^p\times\mathbb{R}^p\to\mathbb{R}, \quad \varepsilon_i\overset{\mathrm{IID}}{\sim}N(0,\sigma_\varepsilon^2), \quad f\perp\!\!\!\perp\{\varepsilon_i\}.$$

Table 11 provides results for GPR-linear, with covariance kernel

$$K(x,x')=\lambda(x-\bar{x})^{\mathrm{T}}(x'-\bar{x}), \qquad \lambda\geqslant 0,$$

and for GPR-FBM-γ, with covariance kernel

Table 11. Misclassification rates for eight data sets: GPR *versus* RP ensembles

Method	*Rates (%) for eye state data*			*Rates (%) for ionosphere data*		
	$n=50$	$n=200$	$n=1000$	$n=50$	$n=100$	$n=200$
RP5-LDA	$42.1_{0.38}$	$38.6_{0.29}$	$36.3_{0.21}$	$13.1_{0.38}$	$10.8_{0.25}$	$9.8_{0.26}$
RP5-QDA	$39.0_{0.39}$	$32.4_{0.42}$	$30.9_{0.87}$	$8.1_{0.37}$	$6.2_{0.37}$	$5.2_{0.20}$
RP5-*k*nn	$39.4_{0.39}$	$26.9_{0.27}$	$13.5_{0.19}$	$13.1_{0.46}$	$7.4_{0.25}$	$5.4_{0.19}$
RP5-GPR-linear	—†	—†	—†	$36.0_{0.11}$	$35.9_{0.16}$	$35.5_{0.28}$
RP5-GPR-FBM-1/2	—†	—†	—†	$36.2_{0.10}$	$35.7_{0.16}$	$35.8_{0.29}$
GPR-linear	$46.6_{0.92}$	$42.3_{0.95}$	$37.5_{0.48}$	$17.3_{0.17}$	$14.6_{0.14}$	$13.5_{0.18}$
GPR-FBM-1/2	$37.0_{0.27}$	$24.0_{0.13}$	$10.3_{0.08}$	$11.2_{0.14}$	$8.0_{0.17}$	$6.3_{0.10}$
GPR-FBM-$\hat{\gamma}$				$12.1_{0.32}$	$8.6_{0.26}$	$6.4_{0.20}$
	Rates (%) for mice data			*Rates (%) for hill–valley data*		
	$n=200$	$n=500$	$n=1000$	$n=100$	$n=200$	$n=500$
RP5-LDA	$25.2_{0.30}$	$23.6_{0.26}$	$23.4_{0.49}$	$36.8_{0.84}$	$36.5_{0.85}$	$32.6_{1.06}$
RP5-QDA	$18.2_{0.29}$	$16.1_{0.24}$	$15.4_{0.45}$	$44.4_{0.34}$	$43.6_{0.31}$	$41.1_{0.33}$
RP5-*k*nn	$11.2_{0.29}$	$2.2_{0.10}$	$0.6_{0.09}$	$49.1_{0.24}$	$47.3_{0.26}$	$36.4_{0.29}$
RP5-GPR-linear	—†	—†	—†	—†	—†	—†
RP5-GPR-FBM-1/2	—†	—†	—†	—†	—†	—†
GPR-linear	$6.5_{0.08}$	$4.5_{0.09}$	$3.8_{0.11}$	$50.2_{0.14}$	$50.0_{0.20}$	$48.5_{0.59}$
GPR-FBM-1/2	$6.6_{0.08}$	$1.2_{0.08}$	$0.1_{0.05}$	$45.3_{0.09}$	$49.8_{0.08}$	$50.7_{0.12}$
GPR-FBM-$\hat{\gamma}$‡	$1.0_{0.11}$	$0.0_{0.00}$		$45.0_{0.09}$	$49.7_{0.10}$	$50.7_{0.13}$
	Rates (%) for musk data			*Rates (%) for arrhythmia data*		
	$n=100$	$n=200$	$n=500$	$n=50$	$n=100$	$n=200$
RP5-LDA	$14.6_{0.31}$	$12.2_{0.23}$	$10.2_{0.15}$	$33.2_{0.42}$	$30.2_{0.35}$	$27.5_{0.30}$
RP5-QDA	$12.1_{0.27}$	$9.9_{0.18}$	$8.6_{0.13}$	$30.5_{0.33}$	$28.3_{0.26}$	$26.3_{0.28}$
RP5-*k*nn	$11.8_{0.27}$	$9.7_{0.21}$	$8.0_{0.15}$	$33.5_{0.40}$	$30.2_{0.33}$	$27.1_{0.31}$
RP5-GPR-linear	$15.2_{0.11}$	$15.5_{0.09}$	$15.5_{0.09}$	$47.3_{0.32}$	$47.5_{0.40}$	$46.2_{0.33}$
RP5-GPR-FBM-1/2	$15.3_{0.10}$	$15.5_{0.10}$	$15.2_{0.11}$	$47.1_{0.33}$	$46.7_{0.27}$	$46.4_{0.25}$
GPR-linear	$15.4_{0.04}$	$11.3_{0.10}$	$9.1_{0.09}$	$38.8_{0.28}$	$33.4_{0.17}$	$27.4_{0.24}$
GPR-FBM-1/2	$9.5_{0.07}$	$7.9_{0.06}$	$5.6_{0.06}$	$34.2_{0.24}$	$29.6_{0.12}$	$26.5_{0.22}$
GPR-FBM-$\hat{\gamma}$				$28.4_{0.27}$	$25.2_{0.22}$	
	Rates (%) for activity recognition data			*Rates (%) for Gisette data*		
	$n=50$	$n=200$	$n=1000$	$n=50$	$n=200$	$n=1000$
RP5-LDA	$0.18_{0.02}$	$0.10_{0.01}$	$0.01_{0.00}$	$15.8_{0.41}$	$10.6_{0.17}$	$9.4_{0.15}$
RP5-QDA	$0.15_{0.02}$	$0.09_{0.01}$	$0.00_{0.00}$	$15.5_{0.40}$	$10.5_{0.19}$	$9.4_{0.16}$
RP5-*k*nn	$0.21_{0.02}$	$0.11_{0.01}$	$0.01_{0.00}$	$16.0_{0.46}$	$11.1_{0.17}$	$9.6_{0.16}$
RP5-GPR-linear	—†	—†	—†	—†	—†	—†
RP5-GPR-FBM-1/2	—†	—†	—†	—†	—†	—†
GPR-linear	$0.05_{0.00}$	$0.00_{0.00}$	0_{0}	$12.4_{0.09}$	$6.8_{0.05}$	$4.5_{0.08}$
GPR-FBM-1/2	$0.15_{0.00}$	$0.03_{0.00}$	$0.00_{0.00}$	$14.0_{0.13}$	$7.0_{0.05}$	$4.5_{0.09}$
GPR-FBM-$\hat{\gamma}$	—	—	—	—	—	—

†Results are currently unavailable. $\hat{\gamma}$ is the maximum likelihood estimator of γ.
‡Records with missing values removed from the data (GPR-FBM-$\hat{\gamma}$ for the mice data).

$$K_\gamma(x,x') = \tilde{K}_\gamma(x,x') - \frac{1}{n}\sum_{j=1}^{n} \tilde{K}_\gamma(x,x_j) - \frac{1}{n}\sum_{i=1}^{n} \tilde{K}_\gamma(x_i,x') + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{K}_\gamma(x_i,x_j), \qquad \gamma \in (0,1),$$

where

$$\tilde{K}_\gamma(x,x') = \frac{\lambda}{2}(\|x\|^{2\gamma} + \|x'\|^{2\gamma} - \|x-x'\|^{2\gamma}), \qquad \lambda \geqslant 0.$$

Scale parameter λ was estimated by using a modified maximum likelihood method. We have omitted GPR-FBM-0.99999 from Table 11, which gives results competitive with the RP ensembles for the hill–valley data. Code for replication of some of the results is provided at `https://haziqjamil.github.io/rec-jrss-reply/` (other results were obtained with Mathematica code).

Xin Bing and Marten Wegkamp (*Cornell University, Ithaca*)

We congratulate Cannings and Samworth on their inspiring paper.

Variable selection

The paper considers mainly projections onto lower d-dimensional spaces with d fixed, though the authors propose to select d from $\mathcal{D} \subseteq \{1,2,\ldots,p\}$ by minimizing the empirical risk when $\mathcal{D}$ is small. But such a minimization depends on $\hat{\alpha}$, that, in turn, should depend on d. Therefore, a potential alternating minimization between α and d might be useful for further improvement. In general, however, ranging over all $d \in \{1,\ldots,p\}$ is not computationally feasible. This is regrettable, since the use of plug-in classifiers, which is advocated in this paper, is problematic in high dimensional settings, where finding suitable hyperplanes $\{x : \theta^{\mathrm{T}} x \geqslant 0\}$ instead is typically preferred. For instance, we can utilize a convex loss l to minimize the penalized empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} l(\theta^{\mathrm{T}} X_i Y_i) + \lambda\|\theta\|_1$$

over $\theta \in \mathbb{R}^p$ and the lasso penalty promotes sparsity in θ by adaptively finding a lower dimensional space. In a way, minimizing

$$\frac{1}{n}\sum_{i=1}^{n} l\{\theta^{\mathrm{T}}(AX_i)Y_i\}$$

over both θ and A could be viewed as trying to achieve the same thing, but it is computationally much more difficult.

Interpretation of threshold α

The random-projection classifier defined in the paper can be viewed as a plug-in rule of the Bayes classifier. Rewriting

$$\mathbb{1}\{\nu_n(x) \geqslant \alpha\} = \mathbb{1}\{\nu_n(x) + (\tfrac{1}{2} - \alpha) \geqslant \tfrac{1}{2}\},$$

we can view $\frac{1}{2} - \alpha$ as an albeit constant bias correction of ν_n due to inherent bias of projecting onto a lower d-dimensional space. It is not difficult to see that

$$\begin{aligned} R(C_n^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) &\leqslant 2\,\mathbb{E}[|\nu_n(X) + \tfrac{1}{2} - \alpha - \eta(X)|] \\ &\leqslant 2\sqrt{\mathbb{E}[|\nu_n(X) - \mu_n(X)|^2]} + 2\,\mathbb{E}[|\mu_n(X) + \tfrac{1}{2} - \alpha - \eta(x)|] \\ &\leqslant \frac{2}{\sqrt{B_1}} + 2\,\mathbb{E}[|\mu_n(X) + \tfrac{1}{2} - \alpha - \eta(X)|]; \end{aligned}$$

see Devroye *et al.* (1996) for the first inequality. The role of B_1 is immediately transparent, as well as the need for a good fit for η and the advantage of using a cut-off α in lieu of $\frac{1}{2}$. Using the same reasoning as in Herbei and Wegkamp (2006) we can derive a more refined result:

$$\begin{aligned} R(C_n^{\mathrm{RP}}) - R(C^{\mathrm{Bayes}}) &\leqslant \inf_{\delta>0}[\mathbb{P}\{|\nu_n(X) + \tfrac{1}{2} - \alpha - \eta(X)| \geqslant \delta\} + 2\delta\,\mathbb{P}\{|\eta(X) - \tfrac{1}{2}| \leqslant \delta\}] \\ &\leqslant \inf_{\delta>0}[\mathbb{P}\{|\nu_n(X) - \mu_n(X)| \geqslant \delta/2\} + \mathbb{P}\{|\mu_n(X) + \tfrac{1}{2} - \alpha - \eta(X)| \geqslant \delta/2\} \\ &\quad + 2\,\delta\,\mathbb{P}\{|\eta(X) - \tfrac{1}{2}| \leqslant \delta\}]. \end{aligned}$$

The first term on the right can be bounded by $\exp(-cB_1\delta^2)$ from Hoeffding's inequality for a constant c as we look at the deviation of an average ν_n around its mean of independently and identically distributed bounded random variables. The second term expresses the need for $\mu_n(X) + \frac{1}{2} - \alpha$ to provide a good bound for $\eta(X)$, whereas the last term reflects the intrinsic difficulty of the classification problem and can be dealt with by a margin condition (Tsybakov, 2004). Again, considering a general cut-off α not necessarily equal to $\frac{1}{2}$ offers additional flexibility to correct the bias and can be viewed as the first-order bias correction for the estimate of $\eta(X)$. This approach circumvents the problematic assumption 2.

Rico Blaser and Piotr Fryzlewicz (*London School of Economics and Political Science*)
We congratulate Cannings and Samworth on their thought-provoking and well-written paper in this promising area of research.

In our work on random-rotation ensembles (Blaser and Fryzlewicz, 2016) we randomly rotated the feature space before applying a high dimensional classifier. In the present paper, the high dimensional feature space is projected into a lower dimensional space before applying a low dimensional classifier. The two strategies are closely related.

In the current paper, projections are performed randomly under the Haar measure. Interestingly, a random rotation followed by a random axis-aligned projection in which d-of-p features are retained is identical to the random projection described in the paper. Our tree-based ensemble classifiers perform axis-aligned projections after rotation and thus effectively describe a random-projection ensemble, whereby the final classification is restricted to a tree-based model.

More generally, we believe decoupling rotation from dimension reduction and dimension reduction from classification is desirable. In particular, such a decomposition addresses the question, if the benefit of a particular random projection arises from an advantageous viewpoint at the problem due to the rotation or from an effective dimension reduction due to the feature selection, as the two operations can be analysed and optimized separately.

The authors also provide interesting insights on the selection of retained projections and the determination of the voting threshold. They note that most random projections are unhelpful in classification: a pattern that we have also observed for random rotations. Hence, a natural question to ask is how we can identify (or explicitly generate) only the most helpful projections.

One way to address this issue is by performing a large number of candidate projections and retaining only the most successful candidates. The authors of the present paper recommend retaining 2% of the generated projections by default: substantially fewer than the 90% of rotations we examined. More accuracy is achieved at the expense of a higher overhead. Alternatively, analytical methods such as principal component analysis can be used to determine successful rotations. In Rodriguez *et al.* (2006), this approach is used for random subsets of the features.

Different subsets of the data frequently benefit from different rotations; for non-linear decision boundaries this is quite evident. Hence, it might also be useful to construct a classifier that rotates different sections of the data independently.

The data-driven selection of the voting threshold suggested by the authors is insightful but is not straightforward to generalize to multiclass problems.

Miguel de Carvalho (*University of Edinburgh*) **and Garritt L. Page and Bradley Barney** (*Brigham Young University, Provo*)
We congratulate Cannings and Samworth for proposing a sturdy method based on randomly compressing feature vectors before classification. Below, we focus on connecting the random-projection ensemble classifier with ideas and concepts from *compressed classification* and *compressed regression* methods. Let $\mathcal{A}=\{A\in\mathbb{R}^{d\times p}: AA^{\mathrm{T}}=I_{d\times d}\}$ be the so-called Stiefel manifold. Similarly to Page *et al.* (2013), Cannings and Samworth first compress the covariates by using projection matrices, but a key difference is that Cannings and Samworth consider a set of independent projections, $A_1,\ldots,A_{B_1}\in\mathcal{A}$, whereas in Page *et al.* (2013) a single projection matrix $A\in\mathcal{A}$ is considered—and treated as a Bayesian parameter. In particular, Page *et al.* (2013) considered a non-parametric Bayesian approach which leads to a principal subspace classifier for a setting similar to that in the current paper and assigns to A a (conjugate) von Mises–Fisher prior distribution on the Steifel manifold. In an analogy to the authors' claim that

> 'in a similar spirit to subsampling and bootstrap sampling, we can can think of each random projection as a perturbation of the original data',

the compressing paradigms described above—based on a single but random A—keep the data as fixed, and posterior sampling about good directions along which to project the data is itself target. Both compressing principles (single $A\in\mathcal{A}$ as a Bayesian parameter, *versus* an ensemble of random $A_1,\ldots,A_{B_1}\in\mathcal{A}$) seem to have their own merits, and we wonder whether the authors could comment on this remark. On another note, the recently proposed compressed regression approach by Guhaniyogi and Dunson (2015) is even closer to the authors' proposal, in the sense that it projects data into an ensemble of directions and uses model averaging to arrive at a final regression model. The focus of Guhaniyogi and Dunson (2015) is on regression itself though, but we also wonder about the authors' view on this. Finally, the practitioner could be left with the question: 'How likely is it for the ensemble classifier to improve over the base classifier on the original feature vectors?'.

Roberto Casarin and Lorenzo Frattarolo (*University Ca' Foscari of Venice*) **and Luca Rossini** (*University Ca' Foscari of Venice and Free University of Bozen-Bolzano*)

Cannings and Samworth are to be congratulated on their excellent research, which has culminated in the development of a characterization of the approximation errors in random-projection methods when applied to classification. We believe that the approach can find many applications in economics such as credit scoring (e.g. Altman (1968)) and can be extended to more general types of classifiers. In this discussion we would like to draw the authors' attention to copula-based discriminant analysis (Han *et al.*, 2013; He *et al.*, 2016).

We consider $X|Y=r$ distributed as a p-dimensional meta-Gaussian distribution and $S|Y=r\sim\mathcal{N}_p(0,\Sigma_r)$, where Σ_r is the linear correlation between variables. Given a $p\times d$ random projection A, $AS|Y=r\sim\mathcal{N}_d(0,\Sigma_r^A)$, where $\Sigma_r^A=A\Sigma_rA^{\mathrm{T}}$. If we assume that the information in the marginals is not relevant for the classification, the Bayes decision boundary depends only on the transformed variables $s_i=\Phi^{-1}\{F(x_i)\}$ with Φ and F the univariate normal and the marginal cumulative distribution functions respectively (Fang *et al.*, 2002), s_i and x_i the ith element of s and x, and the correlation of the two groups

$$\Delta(s;\pi_0,\Sigma_0,\Sigma_1)=\log\left(\frac{\pi_1}{\pi_0}\right)-\frac{1}{2}\log\left\{\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right\}-\frac{1}{2}s^{\mathrm{T}}(\Sigma_1^{-1}-\Sigma_0^{-1})s. \tag{46}$$

Analogously the classifier in the random-projection ensemble will depend only on the random projection of the transformed variables and their covariances. We use the empirical distribution function to obtain the sample version of the transformed variables $S_i=(S_{1i},\ldots,S_{pi})$, with

$$S_{ji}=\Phi^{-1}\left(\frac{1}{n+1}\sum_{k=1}^{n}\mathbb{1}_{\{X_{jk}\leqslant X_{ji}\}}\right),\qquad i=1,\ldots,n,\qquad j=1,\ldots,p. \tag{47}$$

The estimator of Σ_r^A is obtained by maximizing the pseudolikelihood:

$$\hat{\Sigma}_r^A=\frac{1}{n}\sum_{i=1}^{n}AS_iS_i^{\mathrm{T}}A^{\mathrm{T}}\mathbb{1}_{\{Y_i^A=r\}}\qquad\text{for } r=0,1$$

where the asymptotic normality is guaranteed by results in Genest *et al.* (1995) and recently in Segers *et al.* (2014). We propose the following robust quadratic discriminant analysis random-projection ensemble classifier:

$$C_n^{\text{A-RQDA}}(s):=\begin{cases}1 & \Delta(s;\hat{\pi}_0,\hat{\Sigma}_0^A,\hat{\Sigma}_1^A)\geqslant 0,\\ 0 & \text{otherwise.}\end{cases} \tag{48}$$

We are very pleased to thank the authors for their work.

Emre Demirkaya and Jinchi Lv (*University of Southern California, Los Angeles*)

Dr Cannings and Professor Samworth are to be congratulated for their innovative and valuable contribution to the important problem of high dimensional classification. Dimension reduction plays a key role in high dimensional classification, enabling the enhancement of both statistical efficiency and scalability (Fan and Fan, 2008). Through a simple yet ingenious two-level design of using random projections, Cannings and Samworth achieved these goals by proposing the general framework of random-projection ensemble classification with an elegant theory to deal with high dimensionality and to boost the power of existing classification procedures. The general philosophy of random-projection ensemble learning laid out in the paper can also be applicable to many other statistical learning tasks such as clustering and regression.

Our discussion will focus on the perspective of interaction network learning. Understanding large-scale interaction network structures among features can be of fundamental importance in many scientific studies. The problem of interaction network learning has received growing recent interest (Hall and Xue, 2014; Jiang and Liu, 2014; Fan *et al.*, 2015; Kong *et al.*, 2016). Recently, Fan *et al.* (2015)

Table 12. Percentages of retaining all important interactions in model 1 of Fan *et al.* (2015) by RAPID over various settings based on 100 replications when the threshold is chosen to be $[cn/\log(n)]$ following the suggestion in Fan and Lv (2008) with $n=100$ the sample size for each class

c	*Results (%) for p = 100*		*Results (%) for p = 500*	
	d = 5	*d = 10*	*d = 5*	*d = 10*
0.5	0.99	0.97	0.93	0.95
1	1	0.97	0.97	0.95

introduced innovated interaction screening for high dimensional non-linear classification which depends on large precision matrix estimation. An interesting question is whether we can avoid estimating large precision matrices (Fan and Lv, 2016). To provide a partial answer to this question, we borrow the idea in the current paper and suggest a possible extension called random-projection interaction delineation (RAPID).

To illustrate the idea of RAPID, we adopt the framework in Fan *et al.* (2015) and consider a two-class Gaussian classification problem with heterogeneous precision matrices. In view of the Bayes rule, important interactions correspond to non-zero entries of precision matrix difference Ω. RAPID starts by randomly projecting p-dimensional feature vectors to low dimensions d and building classifiers with quadratic discriminant analysis following Cannings and Samworth. Each selected random projection returns a $d \times d$ symmetric matrix from the quadratic form, which can be lifted back to the original p dimensions through the given random projection. Each of B_1 such matrices can be used as a proxy for the original Ω. RAPID then evaluates the significance of each entry by using the t-statistics and ranks the interactions by the magnitude of these t-statistics. A simulation study shows that RAPID can enjoy a nice sure screening property (Fan and Lv, 2008) for interaction screening; see Table 12 for details. It would be interesting to investigate the theoretical properties of this and further extensions.

Josh Derenski, Yingying Fan and Gareth M. James (*University of Southern California, Los Angeles*)
Cannings and Samworth propose a method of classification involving many random projections of the data onto a lower dimensional space and then utilize a base classifier on the projected data to build an ensemble classification rule. They develop theoretical results involving arbitrary base classifiers and highlight the results when applied to particular base classifiers. In addition, they demonstrate the method's strong prediction accuracy with examples involving artificially generated data, and others involving real data.

The random-projection ensemble classifier may also be useful in determining the relative importance of the covariates. The authors suggest that the projections provide weights that can be used as a metric for determining the relative importance of variables. In a similar spirit, using sparse random projections may also assist in determining variable importance. Indeed, after the matrices have been generated and those that yield the smallest test error have been chosen, a variable is selected if the corresponding entries in the selected projection matrices are non-zero. The importance of a variable can be measured by, say, the frequency of the variable being selected.

The authors' proposed method has the flavour of a bagging algorithm, where the data are randomly sampled, a classifier is applied to each new data set and the results are averaged at the end. Hence, it is possible that prediction accuracy could be improved by applying a boosting-type approach. For example, rather than applying the same classifier to each random permutation, one could reweight the observations at each stage, placing higher weight on observations that were misclassified at the previous iteration. This would be somewhat analogous to standard boosting and would potentially provide a similar level of improvement in classification accuracy to that which boosting often has over bagging. Taking this theme one step further, one could choose the random projection conditionally on the performance of the classification method on the previous projection of the data, and then aggregate the results as in boosting.

The extensions suggested above also enable studying the random-projection ensemble classifier under different methodologies for choosing the projection matrices. The authors suggest the possibility of choosing these matrices under different regimes, and there might not be a universally optimal way for selecting these matrices. For example, one sampling scheme might perform better when the goal is inference and another when the goal is prediction. Both these extensions enable the study of this possibility.

Robert J. Durrant (*University of Waikato, Hamilton*)
I thank Cannings and Samworth for an interesting paper, which I am sure will be of interest not only to statisticians but also to researchers in communities such as machine learning.

This paper is initially motivated by the Johnson–Lindenstrauss lemma (JLL), which gives high probability guarantees for the approximate preservation of Euclidean geometry of randomly projected data in $\mathbb{R}^d$ compared with the original data in the embedding space $\mathbb{R}^p$, $d \ll p$. Here I shall discuss some apparent implications of the JLL on the rejection sampling scheme for projection matrices described in this paper. In particular, it is my experience with random projection that for (linear) classification centring and normalizing a set of observations is usually a sensible preprocessing step to apply before random projection, and it appears that may be worthwhile here also. Below follows some informal argument supporting this view.

First note that projection using a sub-Gaussian random-projection matrix implies not only an ϵ–2δ guarantee on norm preservation, but also an ϵ–2δ guarantee on dot product preservation, i.e., under the same conditions as the JLL, for any $\epsilon, \delta \in (0, 1]$ with probability at least $1 - 2\delta$ over the random draws of $A \in \mathbb{R}^{d \times p}$ where the A_{ij} are independently and identically distributed sub-Gaussian with mean 0 and variance σ_A^2 it holds that

$$d\sigma_A^2 \cdot (v^{\mathrm{T}} w - \epsilon \|v\| \|w\|) \leqslant v^{\mathrm{T}} A^{\mathrm{T}} A w \leqslant d\sigma_A^2 \cdot (v^{\mathrm{T}} w + \epsilon \|v\| \|w\|).$$

For any fixed $v, w \in \mathbb{R}^p$ and random A this confidence interval depends on the Euclidean norms $\|v\|$ and $\|w\|$ independently of the angle between these vectors. Thus the JLL implies that, *even for two pairs of vectors with the same angle*, absent normalization, some dot products will be preserved better than others.

In particular, instantiating w as an observation and v as any unit norm classifier learned in $\mathbb{R}^p$, we see that observations with large norms are more likely to be classified differently with respect to v following projection to a fixed dimension $d < p$—i.e. by a sign change in the dot product—than those with small norms. Assuming v was reasonably accurate in the first place this means they will largely be *mis*classified for many instances of projection matrix A, and the corresponding projection matrix instances may risk being rejected—not necessarily because they fail to capture meaningful structure in the data (for the classification task)—instead because of systematic issues introduced by our choice of data representation. Thus it seems that it could be a reasonable step to add data normalization before projection to the authors' algorithm as described here.

Jianqing Fan and Ziwei Zhu (*Princeton University*)
We congratulate Dr Cannings and Professor Samworth for such a brilliant and thought-provoking paper. We believe that it will stimulate extensive research on statistical inference based on randomly projected data.

The authors aim to handle the curse of high dimensionality in classification problems through voting among multiple classifiers based on random data sketches. One of the most attractive aspects of their theories is that the excessive risk of the proposed ensemble classifier depends only on the dimension of the projected data d rather than the dimension of the original data p. To achieve this, the theory requires sufficient dimension reduction conditions. This exact low dimensional structure assumption can be sometimes stringent and some relaxations of the condition are welcome.

Besides overcoming the curse of dimensions, we emphasize that random projection is an accurate and efficient way of dimension reduction when data have (approximately) low dimensional structure. For example, consider the rank k approximation of $\mathbf{X} \in \mathbb{R}^{n \times p}$. Let $\mathbf{A} \in \mathbb{R}^{p \times (k+s)}$ be a random matrix with independently and identically distributed standard Gaussian entries and $\mathbf{Q} \in \mathbb{R}^{n \times (k+s)}$ be the orthonormal column basis of $\mathbf{XA}$. As shown in theorem 1 of Halko *et al.* (2011), for any $s > 1$,

$$E\|\mathbf{X} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{X}\|_{\mathrm{op}} \leqslant \left\{ 1 + \frac{4\surd(k+s)}{s-1} \surd \min(n, p) \right\} \sigma_{k+1},$$

where σ_{k+1} is the $(k+1)$th singular value of $\mathbf{X}$. Since σ_{k+1} is the theoretical minimum of rank k approximation error, this result implies that the column space of the random sketch $\mathbf{XA}$ can capture the top k left singular space of $\mathbf{X}$. It will thus be interesting to investigate the $\sin(\Theta)$ distance between the column space of $\mathbf{Q}$ and the top k left singular space of $\mathbf{X}$. Furthermore, suppose that we create B_1 independent sketches

$\{\mathbf{XA}^{(i)}\}_{i=1}^{B_1}$ as in random-ensemble classification and derive the corresponding column basis $\{\mathbf{Q}^{(i)}\}_{i=1}^{B_1}$. Can we construct an aggregated column basis $\tilde{\mathbf{Q}}$ from $\{\mathbf{Q}^{(i)}\}_{i=1}^{B_1}$ such that $\tilde{\mathbf{Q}}$ will converge to the top k left singular space of $\mathbf{X}$ as B_1 increases to ∞?

Finally, we further stress an important advantage of the ensemble classifier proposed: its full adaptivity to the distributed computing architecture. To implement the random-projection ensemble classification in a distributed computing system, we first let each node computer solve for classification on randomly projected data. Then, according to the estimated risk of the base classifier on each node computer, we can screen out the good projections as described in Section 3 of the paper and construct the final ensemble classifier. Note that the algorithm does not require high communication cost since random projections are small.

Yang Feng (*Columbia University, New York*)
I congratulate Dr Cannings and Professor Samworth on their novel and stimulating contributions to classification using random-projection ensembles (RPEs). It is quite a general framework and we expect to see many follow-up works on the idea combined with some popular classifiers.

Regarding the choice of B_2, the authors did a careful theoretical analysis through assumption 2 and theorem 3. In assumption 2, I wonder whether β should depend on the sample size n or whether the authors believe that there is a universal β for all n. If β in fact turns out to decrease as n increases, we would need to conduct a more delicate analysis regarding the implications on the results of theorem 3 as $n \to \infty$.

Here, I propose a variant of the RPE approach. In this variant, the random projections are not generated independently; instead, the selected B_1 random projections are chosen sequentially and designed to be mutually orthogonal. The intuition is that, by making the random projections mutually orthogonal, the additional contribution of the newly recruited projections could be more significant than those without such constraints. I expect the variant to have a competitive performance when B_1 is small and the problem is high dimensional. A detailed modification is outlined as follows.

First, generate $\mathbf{A}_1$ the same way as the RPE. Now, suppose that we have found the projections $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$, for some k. Then combine the corresponding random projections into the matrix $\mathbf{P}_k = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k)_{p\times(dk)}$. To search for $\mathbf{A}_{k+1}$, first generate B_2 random projections $\{\tilde{\mathbf{A}}_{k+1,b_2}\}_{b_2=1}^{B_2}$ according to the Haar measure on $\mathcal{A}$, and then define $\mathbf{A}_{k+1,b_2} = (I - \mathbf{P}_k(\mathbf{P}_k^{\mathrm{T}}\mathbf{P}_k)^{-1}\mathbf{P}_k^{\mathrm{T}})\tilde{\mathbf{A}}_{k+1,b_2}$ as the orthogonal projection of $\tilde{\mathbf{A}}_{k+1,b_2}$ onto the space $\mathbf{P}_k^{\perp}$, which is the orthogonal complement of $\mathbf{P}_k$. Afterwards, we can follow the same procedure to find the optimal $\mathbf{A}_{k+1}$ by using the new random-projections candidates. At the ensemble step, I propose to use a weighted voting scheme based on the error rate on the test data $\{\mathrm{err}_{b_1}\}_{b_1=1}^{B_1}$ as follows:

$$v_n(x) := \frac{\sum_{b_1=1}^{B_1} w_{b_1} I\{C_n^{\mathbf{A}_{b_1}}(x) = 1\}}{\sum_{b_1=1}^{B_1} w_{b_1}},$$

where $w_{b_1} = \log\{(1 - \mathrm{err}_{b_1})/\mathrm{err}_{b_1}\}$. The final classifier can be created with a data-driven choice of the threshold α by taking into account the weights.

Michael P. B. Gallaugher and Paul D. McNicholas (*McMaster University, Hamilton*)
We congratulate Cannings and Samworth on a very well-written, enjoyable, and interesting contribution. Data collected today are often high dimensional and effective classification techniques for such data are most welcome. In the simulations and the real data analyses, the authors compare the proposed ensemble classifiers with the respective base classifiers as well as 'state of the art' techniques. We note the absence of mixture discriminant analysis, which was introduced in this self-same journal over 20 years ago (Hastie and Tibshirani, 1996) and subsequently studied by others (e.g. Fraley and Raftery (2002)). More general discriminant analysis techniques could also be considered, where a flexible non-Gaussian density is used for each class (see McNicholas (2016), section 9.2, for some discussion). It may also be interesting to consider discriminant analysis using a mixture of factor analysers model (Ghahramani and Hinton, 1997) or an extension thereof (see McNicholas (2016), chapter 3).

For brevity, we consider only mixture discriminant analysis, where the idea is to allow each class to be modelled by using a Gaussian mixture model. For the eye state data set, we take 10 training–test splits with 1000 observations in the training set, similar to the situation in the '$n = 1000$' column of Table 3. Using mixture discriminant analysis via the `mclust` package (Fraley *et al.*, 2017), we obtained an average misclassification rate, for the observations considered unlabelled, of around 0.18; this is a better result than two of the three random-projection classifiers considered. We also note that the mice data set contains missing values; perhaps the authors could clarify how they deal with these missing values. Also, for the hills-and-valleys example, there are multiple such data sets given in the repository at the University of California, Irvine, and it is not clear which are used (presumably it is a training–test pair either with or without noise).

A final point concerns extending the proposed classifiers to more than two classes. In Section 7, the authors mention this possibility and we wonder whether they have actually used the proposed approaches on more than two classes and, if so, what were their experiences?

Milana Gataric (*University of Cambridge*)

I congratulate Cannings and Samworth on their inspiring work that opens numerous avenues for future research. Below I discuss a possible future direction related to the problem of variable ranking mentioned briefly by the authors in the final section of their paper.

Consider the set of axis-aligned projections, namely

$$\mathcal{A}_d = \{A \in \{0,1\}^{d\times p} : AA^{\mathrm{T}} = I_{d\times d}\}.$$

Although this set restricts the originally considered set of transformations, it is nonetheless attractive for at least two reasons. First, the computational complexity reduces considerably since the multiplication of a matrix with $A \in \mathcal{A}_d$ recasts as the selection of the matrix rows (or columns).

Second, this choice of projections paves the way for feature selection in high dimensional classification by adding an additional aggregation step to the originally proposed screening method. Intuitively, by selecting good axis-aligned projections we are selecting features that contribute the most to classification success. Therefore, if $A_{b_1} \in \mathcal{A}_d$ are selected by the original screening method, we could expect that the aggregation such as

$$\hat{a}_j^* = B_1^{-1} \sum_{b_1=1}^{B_1} \mathbb{1}\{(A_{b_1}^{\mathrm{T}} A_{b_1})_{j,j} = 1\}, \qquad j = 1, \ldots, p,$$

provides a good estimation of the classification power for each feature j. Furthermore, if

$$\hat{J} = \{j \in \{1, \ldots, p\} : \hat{a}_j^* \text{ is among the } s \text{ top elements of } \{\hat{a}_1^*, \ldots, \hat{a}_p^*\}\},$$

and $\hat{A}^* \in \mathcal{A}_s$ has non-zero columns corresponding to indices $\hat{J}$, i.e. $\hat{A}^* \in \mathcal{A}_s$ is such that

$$\hat{A}^*_{\cdot,j} \neq \mathbf{0}_s, \qquad j \in \hat{J},$$

we could potentially estimate the lower dimensional axis-aligned projection A^* of X that explains Y, in case such a projection exists.

In view of this last remark, one might draw insight from the special case $X|\{Y=r\} \sim N_p(\mu_r, \Sigma)$, $r = 1, 2$, with discriminative direction $\beta = \Sigma^{-1}(\mu_2 - \mu_1)$ that is s sparse in the sense that $\mathrm{card}(J) \leqslant s$, where $J = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$. In this case, $A^* \in \mathcal{A}_s$ defined such that $A^*_{\cdot,j} \neq 0$, $j \in J$, is the projection that makes assumption 3 of the paper hold. Therefore, we could expect that $\hat{A}^*$ defined as above is a good estimator for A^*. This is demonstrated by a numerical example in Table 13.

Table 13. Results for the model $X|Y = r \sim N_p(\mu_r, \Sigma)$, $\mu_1 = \mathbf{0}_p$, $\mu_2 = \Sigma\beta$, $\beta = (3, 2, 1, \mathbf{0}_{p-3})^{\mathrm{T}}$ and $\Sigma_{j,k} = 0.5^{|j-k|}$†

p	n	$\Vert A^* - \hat{A}^*\Vert_{\mathrm{F}}$	$(\hat{a}_1^*, \hat{a}_2^*, \hat{a}_3^*, max_{j=4,\ldots,p}\, \hat{a}_j^*)$
200	100	$0_{(0)}$	$(0.4317_{(0.0068)}, 0.3512_{(0.0069)}, 0.1312_{(0.0033)}, 0.0633_{(0.0010)})$
200	200	$0_{(0)}$	$(0.4379_{(0.0067)}, 0.3467_{(0.0067)}, 0.1287_{(0.0023)}, 0.0650_{(0.0007)})$
400	100	$0.05_{(0.0221)}$	$(0.2754_{(0.0029)}, 0.2557_{(0.0028)}, 0.1501_{(0.0012)}, 0.1018_{(0.0013)})$
400	200	$0.03_{(0.0172)}$	$(0.2794_{(0.0028)}, 0.2516_{(0.0028)}, 0.1499_{(0.0012)}, 0.1054_{(0.0010)})$

†Here linear discriminant analysis is used as the base classifier, $d = s = 3$, $B_1 = 500$ and $B_2 = 50$, and the experiment is repeated 200 times.

Table 14. Results for the model corresponding to model 1 of the paper†

p	n	$\|A^* - \hat{A}^*\|_{\mathrm{F}}$	$(\hat{a}_1^*, \hat{a}_2^*, \max_{j=3,\ldots,p} \hat{a}_j^*)$
100	50	$0.22_{(0.0690)}$	$(0.2978_{(0.0028)}, 0.2966_{(0.0031)}, 0.0514_{(0.0039)})$
100	100	$0.16_{(0.0545)}$	$(0.2995_{(0.0028)}, 0.2980_{(0.0028)}, 0.0530_{(0.0041)})$
100	200	$0.20_{(0.0603)}$	$(0.2966_{(0.0035)}, 0.2949_{(0.0036)}, 0.0520_{(0.0041)})$

†Here quadratic discriminant analysis is used as the base classifier, $s=2$, $d=5$, $B_1=500$ and $B_2=150$, and the experiment is repeated 100 times.

Moreover, because of the flexibility of the original proposal in regard to the base classifier, it is to be expected that these concepts translate well to all scenarios with sparse class boundaries, not necessarily linear. In particular, in Table 14, this is illustrated on model 1 of the paper.

Tilmann Gneiting and Sebastian Lerch (*Heidelberg Institute for Theoretical Studies and Karlsruhe Institute of Technology*)

We congratulate Cannings and Samworth on an impressive paper that spans the gamut from theory to computation and empirical studies. The use of projections indeed has a rich history in statistics, with classical work by Friedman and Stuetzle (1981), page 823, and Huber (1985), pages 435 and 499, on projection pursuit alluding to classification in various places.

The paper restricts attention to binary classification under the symmetric 0–1 loss function, under which the Bayes rule assigns class label 1 if the conditional predictive probability thereof exceeds the threshold $\frac{1}{2}$. The associated voting threshold α in the definition of the random-projection ensemble classifier (2) in terms of the ensemble vote (1) is chosen by minimizing empirical 0–1 loss in cross-validation mode.

Applications frequently call for class probabilities so that decision makers can find the Bayes classifier under the loss function at hand, which might be asymmetric (Hand (1997), chapter 8). The class probability setting can be handled similarly, by modelling a non-decreasing calibration function $B:[0,1]\rightarrow[0,1]$ that assigns a calibrated predictive probability to the ensemble vote. The aforementioned threshold α can then be thought of as satisfying $B(\alpha)=\frac{1}{2}$. The calibration function B could be modelled by the cumulative distribution function of the beta family, as proposed by Ranjan and Gneiting (2010), with the beta parameters being estimated by minimizing the empirical loss under a proper scoring rule for probability forecasts of a binary event (Gneiting and Raftery (2007), section 9.1). A more general calibration approach has recently been proposed by Bassetti *et al.* (2017). In any practical setting, competing methods for class probability estimation can be compared with proper scoring rules (Gneiting and Raftery (2007), section 3.2) and Murphy diagrams (Ehm *et al.*, 2016). Theoretically, a natural question is whether the asymptotic results in the paper admit generalizations in these directions.

Whether we seek class probabilities or a classifier, the ensemble vote might average over class probability estimates, as opposed to averaging over classifiers. In conjunction with the k-nearest-neighbour approach for the projected data, such an approach can be implemented straightforwardly. Intuitively, the k-nearest-neighbour class probability estimate carries additional information, compared with the majority vote classifier, so we might ask whether the asymptotic results in Section 4.3 could be sharpened, and empirical results for the RP-kNN$_d$ techniques in Section 6 could be improved, by using class probability estimates as input for the ensemble vote.

Lucas Janson (*Stanford University*)

I congratulate Cannings and Samworth on an excellent paper. The methodological idea is general, intuitive and appealing, and the theoretical analysis and extensive simulations (including substantial supplemental materials) supports its use and aids in its understanding. I draw two connections which may be enlightening and suggestive of future directions.

(a) *Axis-aligned projections*: the authors mention in Section 7 the potential for using axis-aligned projections instead of Haar-distributed projections. If $B_2=1$, this results in a meta-algorithm that is very similar to the random-subspace method (Ho, 1998). Axis-aligned projections also correspond to randomly dropping features, raising the connection to dropout training (Hinton *et al.*, 2012). Dropout can also be viewed as a form of regularization (Wager *et al.*, 2013), as can ensembling random projections (Durrant and Kabán, 2015), although both these examples assume uniformly distributed projections ($B_2 = 1$). It would be interesting, and potentially computationally beneficial, to connect the ensembling in the present paper (with $B_2 > 1$) to some form of regularization.

(b) *Projection pursuit regression*: the authors motivate random projections by the Johnson–Lindenstrauss lemma, pointing out that Haar-distributed projections have nice distance preserving properties. They then advocate doing some selection of the projections to choose those with the highest predictiveness, or ability to distinguish the classes. The result is a trade-off between predictiveness and variability among the projections ($B_2 = 1$ gives little predictiveness and maximal variability, whereas $B_2 \to \infty$ makes B_1 irrelevant, since the most predictive projection will be chosen every time with no variability), with the advantages of balancing this trade-off shown by the theory and simulations. Whereas the approach proposed starts with highly variable projections and then selects for predictiveness, a conceptual alternative is first to select projections with high predictiveness—the goal of projection pursuit regression (see, for example, Friedman and Stuetzle (1981))—and then to add in variability. For instance, one could choose any algorithm for projection pursuit regression but apply it to random subsets of the features and then use the resulting projections for classifications which are then ensembled. The trade-off between predictiveness and variability also makes clear a close connection with random forests, where tree classifiers are randomized by considering random subsets of features at each tree splitting, and then the resulting classifications are ensembled. A possible advantage of random-projection ensemble classification over random forests is that it seems to give the user more flexibility, and indeed one could choose the base classifier to be tree based. I wonder how such an implementation would compare with random forests in the paper's examples where random forests produced the lowest error (simulated model 4 and arrhythmia data).

Dehan Kong (*University of Toronto*)
I congratulate Cannings and Samworth for their thought-provoking and fascinating work on random-projection ensemble classification. They introduce a very general method for high dimensional classification based on a careful combination of the results of applying an arbitrary base classifier to random projections of the feature vectors into a lower dimensional space. The authors show that the test excess risk of the random-projection ensemble classifier can be controlled by terms that do not depend on the original data dimension. This is a very interesting and surprising finding. This work is a substantial contribution to high dimensional classification problems.

I have several comments about the paper. First, it looks like the performance of the method proposed sometimes may improve when p increases, e.g. the misclassification rates of RP-QDA$_5$ and RP-$k$nn$_5$ for model 1. It is unclear why the performance improves because intuitively we would expect higher misclassification rates when p increases. Second, the third term of the error bound in theorem 3 depends on the constant β, which is defined in assumption 2. I guess that this β may depend on the dimension p, although it is unclear what the relationship between β and p is. If it happens that $\beta \to 0$ when $p \to \infty$, B_2 may have to depend on p to make the third term negligible. Third, it might be useful to extend the idea in this paper to high dimensional regression problems, especially the cases when the sparsity assumption does not hold.

Baibing Li (*Loughborough University*) **and Keming Yu** (*Brunel University, Uxbridge*)
The paper introduces a general method for high dimensional problems in discriminant analysis by applying random projections of the feature vectors into a lower dimensional space. Discriminant analysis is usually considered to be supervised learning where the desired output value (group label) of each object in the training sample is known *a priori* (Bishop, 2006). Here we briefly discuss how this random-projection ensemble classification method could be extended to cluster analysis or unsupervised learning where the desired output value of each object is unknown.

Consider the pair (X, Y) taking values in $\mathbb{R}^p \times \{1, -1\}$. Let the corresponding training sample be $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where the group labels y_i $(i = 1, \ldots, n)$ in cluster analysis are assumed unknown. Li (2006) developed a clustering-function-based method where a linear clustering function $\alpha + \beta^{\mathrm{T}} X$ with coefficients α and β is estimated for clustering purposes through sign eigenanalysis such that each object $X = x$ is classified into one group (labelled 1) or the other (labelled -1), depending on the sign of $\alpha + \beta^{\mathrm{T}} x$, i.e. the classifier is

$$C(x) := \begin{cases} 1 & \text{if } \alpha + \beta^{\mathrm{T}} x \geqslant 0, \\ -1 & \text{otherwise.} \end{cases}$$

We point out that this classifier can be used to deal with $p > n$ clustering problems if random projections can be applied properly. Specifically, for a random $d \times p$ projection matrix A, we define the projected data $z_i^A := Ax_i$ and $y_i^A := y_i$ for $i = 1, \ldots, n$. Then we look for a linear clustering function $\alpha^A + (\beta^A)^{\mathrm{T}} Z^A$ and $y_i^A := y_i$ simultaneously. This linear clustering function can be further extended to a quadratic classifier, i.e. $\alpha^A + (\beta^A)^{\mathrm{T}} Z^A + (Z^A)^{\mathrm{T}} D^A Z^A$, where D^A is a $d \times d$ coefficient matrix to be estimated.

Then we need a criterion function $R(C)$ to choose good random projections. There are many criteria for evaluating classification results. Two widely used criteria are

(a) to maximize the ratio of the between-group to the within-group variance and
(b) to minimize the trace of the within-group covariance matrix.

The former is Fisher's criterion in discriminant analysis and the latter is more commonly used in cluster analysis (Everitt *et al.*, 2011).

For the chosen criterion R_n^A with a projection A, and $B_1, B_2 \in \mathbb{N}$, let $\{A_{b_1,b_2} : b_1 = 1, \ldots, B_1; b_2 = 1, \ldots, B_2\}$ denote independent projections. Following equation (7), for $b_1 = 1, \ldots, B_1$, let

$$b_2^*(b_1) := \mathrm{sarg} \min_{b_2 \in \{1, \ldots, B_2\}} R_n^{A_{b_1,b_2}},$$

and set $A_{b_1} := A_{b_1, b_2^*(b_1)}$. We consider the random-projection classifier using the independent projections $A_1, \ldots, A_{B_1}$. The ultimate assignment of each object, y_i for $i = 1, \ldots, n$, is made by aggregation and a vote.

In summary, we outline a new extension of the Fisher linear discriminant function and quadratic discriminant function to cluster analysis with $p > n$. We shall pursue this research in detail elsewhere.

Yurong Ling, Xiaochen Yang and Jing-Hao Xue (*University College London*)
We congratulate Cannings and Samworth on their elegant contribution to both theory and methodology of ensemble learning. In Section 6.1.3, the random-projection (RP) ensemble classifier with non-linear base classifiers demonstrated encouraging performance when the class boundaries are non-linear. What happens if the non-linearity is taken to extremes?

We experiment on the Swiss roll data, a type of benchmark data in manifold learning. We first generate three-dimensional data of two classes lying on a non-linear manifold as illustrated in Fig. 6, and then

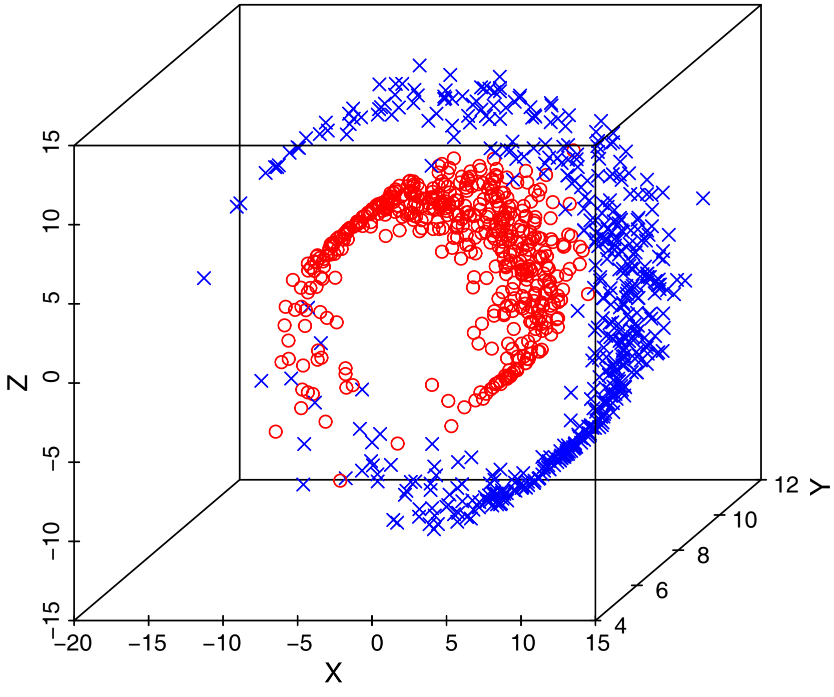


Fig. 6. Three-dimensional Swiss roll data

Table 15. Misclassification rates for the synthetic non-linear data†

Classifier	*Rates (%) for p = 100*			*Rates (%) for p = 1000*		
	n = 50	*n = 200*	*n = 1000*	*n = 50*	*n = 200*	*n = 1000*
RP-$k$nn$_2$	19.21	12.82	9.74	32.65	25.54	21.71
RP-$k$nn$_2$-a	16.77	6.64	3.50	25.14	12.09	8.10
RP-$k$nn$_5$	13.55	6.70	3.80	31.42	22.52	20.36
RP-$k$nn$_5$-a	16.35	7.61	2.44	*15.97*	*5.74*	*3.41*
$k$nn	*7.48*	*3.01*	*1.45*	20.00	11.95	7.94
RF	28.13	8.31	2.53	43.24	20.07	7.29
Radial SVM	49.38	47.04	26.29	50.00	49.97	50.16

†The best misclassification rates are in italics.

augment the data into a much higher dimensional space by adding independent dimensions of $N(5, 1)$ data such that there are many features irrelevant to classification.

In our experiments, we set $B_1 = 500$, $B_2 = 50$, $\pi_1 = 0.5$, $n_{\text{test}} = 1000$ and $N_{\text{reps}} = 100$, and try Gaussian-distributed projections and axis-aligned projections. For the k-nearest-neighbours classifier $k$nn and its RP ensemble versions, we choose k from 1 to 7 via leave-one-out cross-validation; the parameters of random-forests RF and the radial support vector machine SVM are the same as in the paper. The results are listed in Table 15: $k$nn achieves the lowest misclassification rate when $p = 100$, and RP-$k$nn$_5$-a (axis aligned) performs the best when $p = 1000$.

These patterns may be due to two reasons: most features are irrelevant to classification and the class boundary is non-linear. When $p = 1000$, whereas the entire class information is stored in only two features in our experiments, Gaussian projections give weights to all features and hence limit the effectiveness of the RP ensemble; $k$nn does similarly. Therefore, by using axis-aligned projections, which place zero weight on unselected lower dimensions, results have improved considerably. When $p = 100$, the relative effect of irrelevant features is lessened, but the limitation of the non-linear class boundary becomes relatively more obtrusive to linear projections.

Meimei Liu and Guang Cheng (*Purdue University, West Lafayette*)

We congratulate Cannings and Samworth for an inspiring piece of work. Accuracy and stability are two main principles in designing classification algorithms; see Yu (2013). This short note empirically examines how these two measures are affected by the choice of (B_1, B_2, d), which in turn determines the computational cost of the proposed method in the paper.

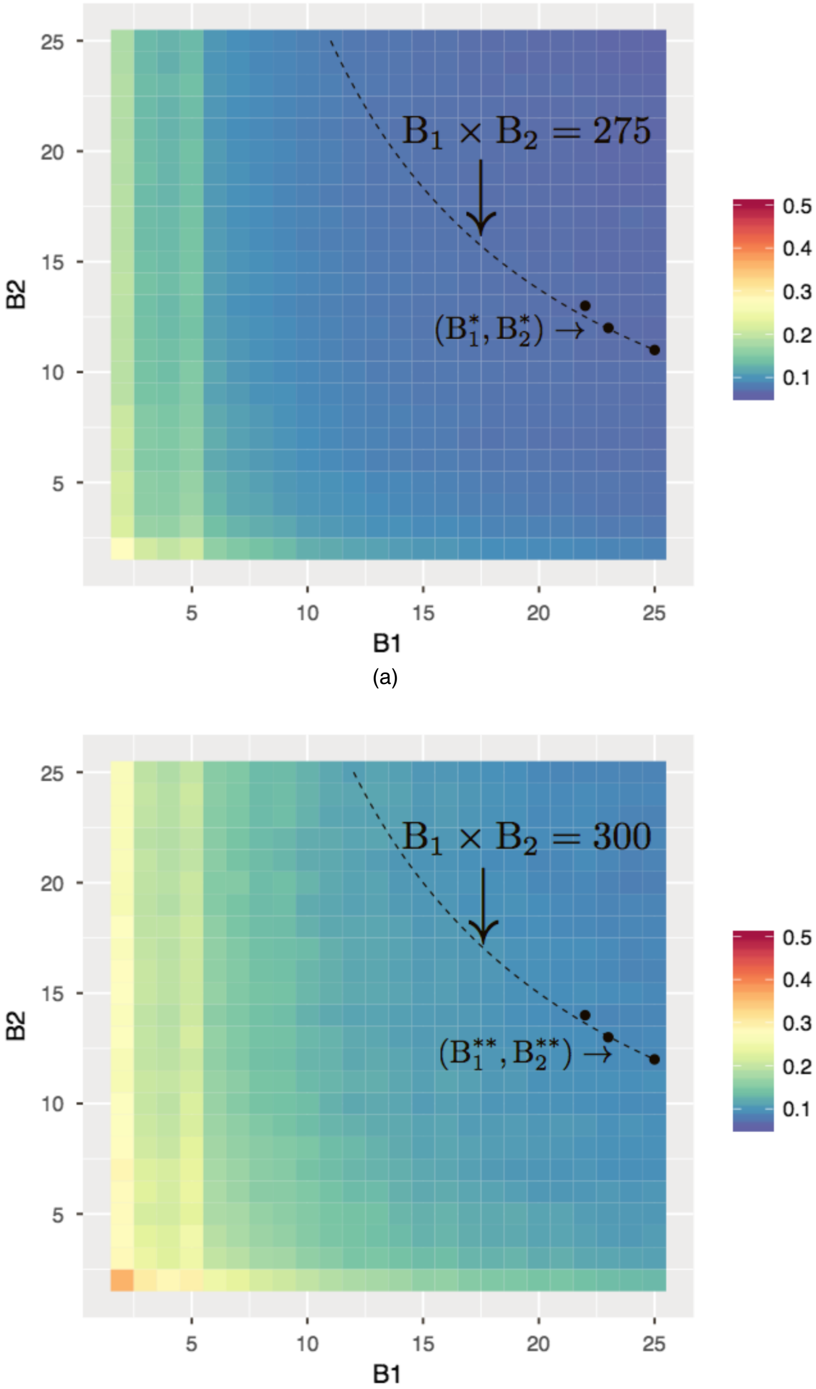
According to Sun *et al.* (2016), one (statistically meaningful) way to define instability for a classification procedure Φ is

$$\text{CIS}(\Phi) = \mathbb{E}_{\mathcal{D}_1, \mathcal{D}_2}[d(\hat{\phi}_{n1}, \hat{\phi}_{n2})] \tag{49}$$

where $d(\hat{\phi}_{n1}, \hat{\phi}_{n2}) = \mathbb{P}_X\{\hat{\phi}_{n1}(X) \neq \hat{\phi}_{n2}(X)\}$ and $\hat{\phi}_{ni} = \Phi(\mathcal{D}_i)$ is the classifier trained on the basis of the sample $\mathcal{D}_i$ for $i = 1, 2$, which is drawn from the same population.

In Fig. 7, we fix $d = 5$ and study how the misclassification rate and CIS are affected by different combinations of (B_1, B_2). Fig. 7(a) shows that once B_1 is sufficiently large the misclassification rate will not change too much as B_2 grows. However, the pattern of misclassification rates is roughly the same as B_1 grows under different choices of B_2. This might indicate that B_1 plays a more prominent role than B_2 in determining the misclassification rate. By examining Fig. 7(b) on CIS in a similar way, we find that the roles of B_1 and B_2 are more comparable, though. We further investigate the least computational cost needed to achieve the best accuracy and stability. In Fig. 7(a), three dots, denoted as (B_1^*, B_2^*), are found to have the smallest value of $B_1 \times B_2$, i.e. 275, among all combinations of (B_1, B_2), leading to the smallest misclassification rate. In contrast in Fig. 7(b) for CIS, we need a higher computational budget, i.e. $B_1^{**} \times B_2^{**} = 300$, to obtain the best stability.

However, the projection dimension d is another factor in determining the computational cost. Hence, in Fig. 8, we fix the total number of random projections, i.e. $B_1 \times B_2$, while varying d. Fig. 8(a) shows that



(a)

(b)

Fig. 7. Heat map of (a) the misclassification rate ($d=5$) and (b) CIS ($d=5$) under various values of (B_1, B_2): the training data set of size 200 and the testing data set of size 1000 were generated following Section 6.1.2 of the paper; the *k*nn classifier is considered here; CIS is calculated by averaging the disagreement of two classifiers on the testing data with 100 replications

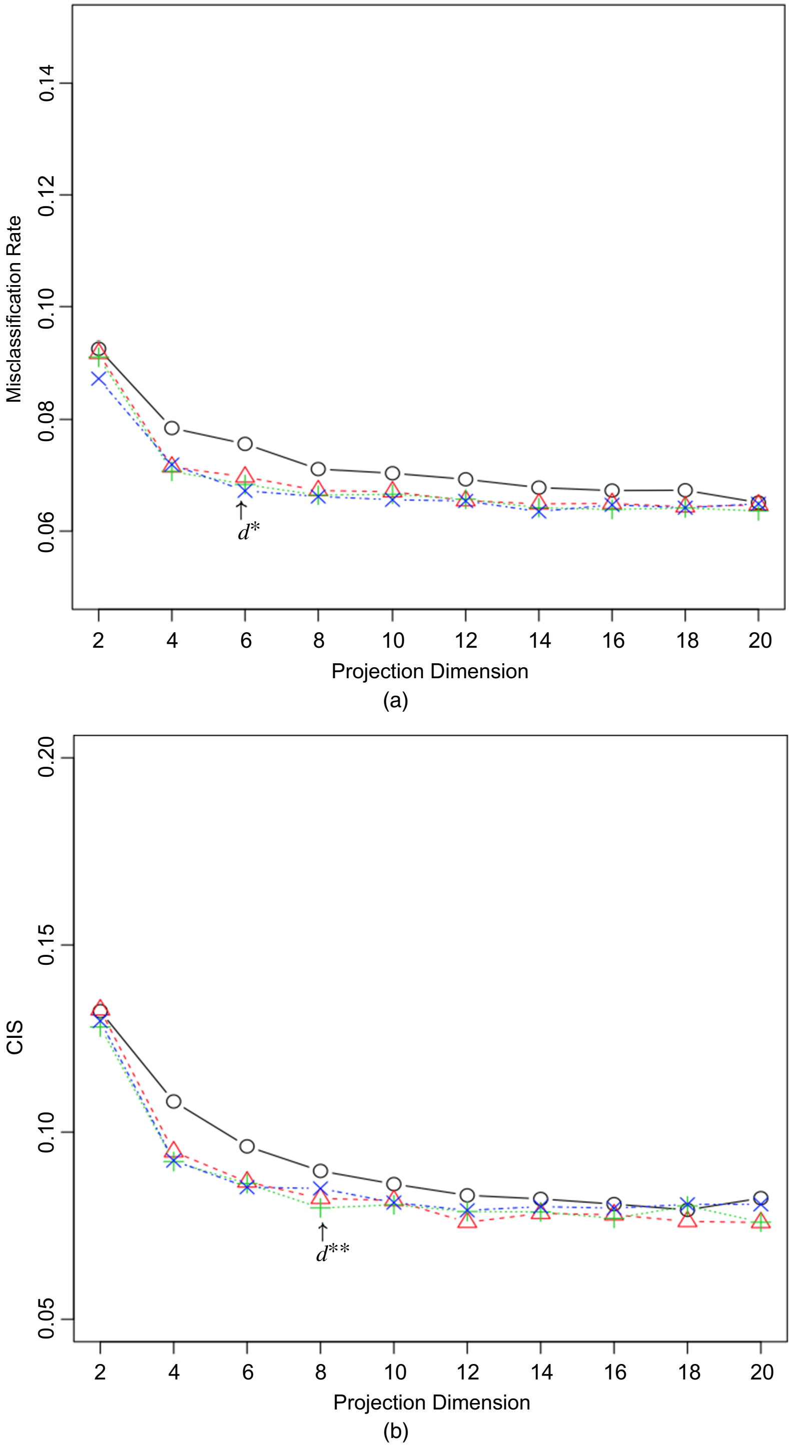


Fig. 8. Effect of projection dimension on (a) the misclassification rate and (b) CIS (d ranges from 2 to 20, $n=200$ and $B_1 \times B_2 = 200$; the $k$nn classifier is considered here): ○, $B_1=20, B_2=10$; △, $B_1=40, B_2=5$; +, $B_1=50, B_2=4$; ×, $B_1=100, B_2=2$

an increase in B_1 leads to a smaller misclassification rate, but this improvement is no longer obvious when B_1 is sufficiently large. However, the pattern for CIS in Fig. 8(b) is not that clear. Another phenomenon in Fig. 8 is that the misclassification rate and CIS in all curves cannot be further improved as d grows beyond some critical point d^* and d^{**} respectively. Since $B_1 \times B_2$ are fixed in all curves, a *sharp* lower bound of d might be viewed as the computational limit of the proposed algorithm from a statistical perspective.

All these empirical observations require new theoretical understanding of high dimensional classification problems from the perspective of computational cost.

Xiaoou Lu and Jing-Hao Xue (*University College London*)
We congratulate Cannings and Samworth on their impressive paper. We suggest that their random-projection (RP) ensemble may be enhanced through regularizing the diversity of base classifiers.

There are emprical results indicating that diversity can be benefical to ensemble learning (Dietterich, 2000; Kuncheva and Whitaker, 2003). Although diversity is not necessarily an issue with RP, lack of diversity may happen with the proposed RP ensemble, as it retains only the projection that yields the smallest estimate of test error in each of B_1 blocks, which may result in B_1 similar base classifiers when the influential components are few, for instance.

For the RP ensemble to consider the trade-off between the diversity and the accuracy of the base classifiers selected, a simple way perhaps is to penalize the similarity of projection matrices A_{b_1,b_2^*}, for $b_1 = 1, \ldots, B_1$, while learning the optimal matrices. A greedy forward strategy is as follows. Suppose that $D(A_{i,b_2}, A_{j,b_2})$ is a measure of dissimilarity between matrices A_{i,b_2} and A_{j,b_2}. When $b_1 = 1$, let

$$b_2^*(b_1) = \underset{b_2 \in \{1,\ldots,B_2\}}{\operatorname{sargmin}} \; R_n^{A_{b_1,b_2}};$$

when $b_1 > 1$, regularize the dissimilarity as

$$b_2^*(b_1) = \underset{b_2 \in \{1,\ldots,B_2\}}{\operatorname{sargmin}} \left\{ R_n^{A_{b_1,b_2}} - \lambda \frac{1}{b_1 - 1} \sum_{j<b_1} D(A_{b_1,b_2}, A_{j,b_2^*(j)}) \right\},$$

with positive trade-off parameter λ to be tuned. The penalty

$$\frac{1}{b_1 - 1} \sum_{j<b_1} D(A_{b_1,b_2}, A_{j,b_2^*(j)})$$

is simply the average dissimilarity between a candidate projection matrix A_{b_1,b_2} and the previously selected projection matrices $A_{j,b_2^*(j)}$. Alternatively we can regularize the diversity of base classifiers' parameters or outputs, as practised in ensemble learning (Yu *et al.*, 2011; Li *et al.*, 2012).

Jorge Mateu (*University Jaume I, Castellón*)
Cannings and Samworth are to be congratulated on a valuable contribution and thought-provoking paper on high dimensional classification focused on applying base classifiers to random projections, This is a timely and extremely interesting topic transversely involved in various areas of science with, as the authors state, a plethora of applications, including spam filtering, fraud detection, medical diagnoses, market research and natural language processing. The random-projection ensemble classifier proposed seems to be competitive, outperforming other widely used in the literature high dimensional classifiers, such as linear discriminant analysis, support vector machines, kernel methods or nearest neighbour classifiers. Our discussion here is more focused on linking the random-projection ensemble classifier with problems related to spatial and functional data.

Consider the problem of detecting features of general shape in d-dimensional point processes in the presence of substantial clutter. In this context, we are interested in removing the clutter to clean and highlight the corresponding features leading to a classification problem. Here the trick is handling the spatial structure as the observations are spatially correlated, and thus the more classical classifiers do not perform quite correctly as they do not properly handle this structure. Alternatively, we can find classification methods based on a stochastic version of the expectation–maximization algorithm that working on local versions of the product density (local indicators of spatial association functions) can provide classification rules that outperform rules based on the kth-nearest-neighbour technique (see, for example, Mateu *et al.* (2007, 2010)). Additionally, local indicators of spatial association functions are density functions associated with individual locations in the d-dimensional space, providing the chance to perform functional data analysis with spatial correlation (see, for example, Bohorquez *et al.* (2016, 2017)). Thus we face a classification problem for functional data.

We note that a model-based clustering technique has its roots in the linear discriminant analysis point of view presented in Section 4.1 of the paper, However, the implicit assumption of Gaussianity prevents its general use in the context of spatial processes. We thus advocate the implementation of a random-projection ensemble classifier adapted to the case of spatial correlation. On the basis of our experience an adapted version of the k-nearest-neighbour classifier in Section 4.3 could do a good job, in particular when the dimension of the process is larger than 2. For planar events, we might additionally think of sample splitting by using a kind of subsampling strategy.

Fionn Murtagh (*University of Huddersfield*) **and Pedro Contreras** (*Thinking Safe, Egham*)
We are grateful for this very important work in the area of supervised classification.

For unsupervised classification, there are certainly major benefits in aggregating over many random projections. There is a clear distinction with regard to the central Johnson–Lindenstrauss lemma, that expresses the precision of low dimensional mapping. Just as in unsupervised classification and related pattern recognition, there is no curse of dimensionality whatsoever given that very high dimensional spaces are naturally and inherently ultrametric, i.e. endowed with hierarchical topology (see Murtagh (2017)). With massive high dimensional spaces being inherently hierarchical, a practical focus of interest becomes (e.g. Murtagh and Contreras (2015, 2017)) the practical scaling of data that furnishes hierarchical clustering (Critchley and Heiser, 1988). Our approach to random-projection-based, linear computational time, hierarchical clustering, must differ in methodology in regard to distribution of the random axes (here uniform, that allow for the compactification of massive data sets), normalization and aggregation to the mean of projections (in Murtagh and Contreras (2015), compared with the dimensionality reduction implementation of Kaski (1998)).

From all of this, there are the practical benefits, for unsupervised classification and related analytical methods, from our outputs that are processed through random projection, just to express in more easily and directly interpretable number systems. See Murtagh (2016).

This led us to take one of the data sets used here to carry out unsupervised clustering. We used the ionosphere data set (Section 6.2.2), of dimensions 351×34, so arriving at random projection (99 random axes, and also a close look at just one random axis), and then converted from decimal (10-ary or 10-adic) to binary (2-adic), with the top level partition read off from the hierarchical tree. For the two predefined classes in this data set, 'good' and 'bad', we determined recall and precision performance measures and then misclassification rates. For the 99 random-projections-based approach, for the first class, there were recall and precision of respectively 72.9% and 77.4%; for the second class, these measures were respectively 61.9% and 56.1%. The misclassification rate overall was poor: 31.1%. For one experiment with just a single use of a random-projection axis, with the overall processing context being identical, we found recall and precision measures for the first class of 91.1% and 75.4% and for the second class 46.8% and 74.7%. For that case, the misclassification rate was 24.8%.

Radka Sabolová (*The Open University, Milton Keynes*) **and Paul Marriott** (*University of Waterloo*)
Thank you for a very stimulating paper. Our comments are twofold, involving joint work with Frank Critchley.

First, as indicated in his oral discussion contribution, there might be potential synergies between random-projection (RP) and sufficient dimension reduction (SDR) methodologies. Recall that the (in principle, checkable) condition $Y \perp\!\!\!\perp X|AX$ defines what it means for span(A^{T}) to be an SDR subspace for $Y|X$ regression. And that, under mild conditions (implicitly assumed), the intersection of all such subspaces is itself an SDR subspace, called the *central subspace* for this regression and denoted $S_{Y|X}$. Possible synergies include the following.

(a) SDR$\rightarrow$RP? The idea here is to gain precision by ignoring any redundant dimensions, perhaps as follows.
   (i) With $Z := AX$, use SDR to estimate $S_{Y|Z}$ as span(A_Z^{T}).
   (ii) Put $W := A_Z Z$, so that dim(W) $\leqslant$ dim(Z).
   (iii) Proceed as in the paper, classifying via W, not Z.

(b) SDR $\leftarrow$ RP? The idea here is to learn about $S_{Y|X}$ via an ensemble of $S_{Y|Z}$ subspaces. There is a variety of ways in which this might be done, involving both random and, for given training data, fixed projections. In the latter case, cognate ideas appear in Wang *et al.* (2016).

(c) SDR $\rightarrow$ RP, again? The idea here is to use an estimate of the central subspace $S_{Y|X}$ to give *interpretable* classifiers, perhaps as follows.

(i) Let SDR ← RP (or any other procedure) estimate $S_{Y|X}$ as span($A_\circ^{\mathrm{T}}$).
(ii) Proceed as in the paper, classifying via the explicit variables in $Z_\circ := A_\circ X$.
(iii) Additionally, for the training data, plotting Y against $Z_\circ$ may itself directly suggest a suitable classifier.

Finally, we note that SDR may detect unexpected subpopulation regression structure (Cook and Critchley, 2000).

Second, random-projection ideas may have an important part to play in the challenging problem of inference in large, sparse, discrete data settings. In particular, Marriott *et al.* (2015, 2016) studied how sampling distributions can be dominated by continuous or discrete aspects. It may transpire that diagnostics for the adequacy of a continuous approximation should be best calculated after a random projection of the underlying sparse simplex.

Chengchun Shi, Rui Song and Wenbin Lu (*North Carolina State University, Raleigh*)
We congratulate Cannings and Samworth for their thoughtful paper on high dimensional classification. Supervised classification is quite a challenging task when the number of features p is comparable with or much larger than the sample size n. In the paper, the authors propose to apply an arbitrary base classifier based on random projections of the feature vectors and to use a data-driven approach to aggregate these results. Specifically, they divide the random projections into non-overlapping blocks, select the projection that gives the smallest estimated test error and aggregate the classifiers on the basis of these selected random projections. By doing so, they show that the test error of their classifier can be controlled by terms that are independent of p (theorem 5).

We note that theorem 5 depends on assumption 2, which requires the distribution function of the estimated test error of a random-projection-based classifier to be close to that of the minimum estimated test error over all random projections. It implicitly assumes that the constants β_0, β and ρ that are involved in that condition are independent of p. When these constants depend on p, however, the upper bound for the test error of their classifier will be dependent on p as well. It would be helpful if the authors could elaborate more on this condition. For example, which values will β_0, β and ρ take if we use linear discriminant analysis, quadratic discriminant analysis or the k-nearest-neighbour classifier as the base classifiers?

The authors provided an upper bound on the test error of their proposed classifier. It would be interesting to study the asymptotic distribution of the test error of their random projection ensemble classifier. In practice, the asymptotic distribution of the test error of the classifiers is often very useful, where a researcher may need it for testing hypotheses or constructing confidence intervals. Statistical inference of the test error of the classifier proposed remains unclear but would be quite useful in expanding the scope of applications. Does the asymptotic distribution of the test error exist or not? If it exists, what are the conditions on n, p, B_1 and B_2? We would appreciate comments from the authors on the possibilities and difficulties in derivations of such inference.

Seung Jun Shin (*Korea University, Seoul*) **and Chaowen Zheng and Yichao Wu** (*North Carolina State University, Raleigh*)
Cannings and Samworth are to be congratuated for an insightful and thought-provoking paper. We would like to add a comment that the idea proposed can be naturally extended to more general problems beyond binary classification.

Although the authors focused on the 0–1 loss, a general loss function such as squared loss for conditional mean regression and check loss for conditional quantile regression can be used. Suppose that we are given a pair of $(\mathbf{X}, Y)$ taking values in $\mathbb{R}^p \times \mathbb{R}$ with joint distribution P. The goal is to find a function $f : \mathbb{R}^p \to \mathbb{R}$ that minimizes

$$R(f) := \int_{\mathbb{R}^p \times \mathbb{R}} L\{f(\mathbf{x}), Y\} \, \mathrm{d}P(x, y), \tag{50}$$

where L denotes an appropriate loss function.

Let f_n denote the minimizer of $R_n(f)$, an empirical version of expression (50). Given a random projection $\mathbf{A}$, we can compute a base regression estimator $f_n^{\mathbf{A}}$ corresponding to f_n by using the projected data. Following the random-projection ensemble (RPE) idea, we set

$$\nu_n(\mathbf{x}) = \frac{1}{B_1} \sum_{b_1=1}^{B_1} f_n^{\mathbf{A}_{b1}}(\mathbf{x})$$

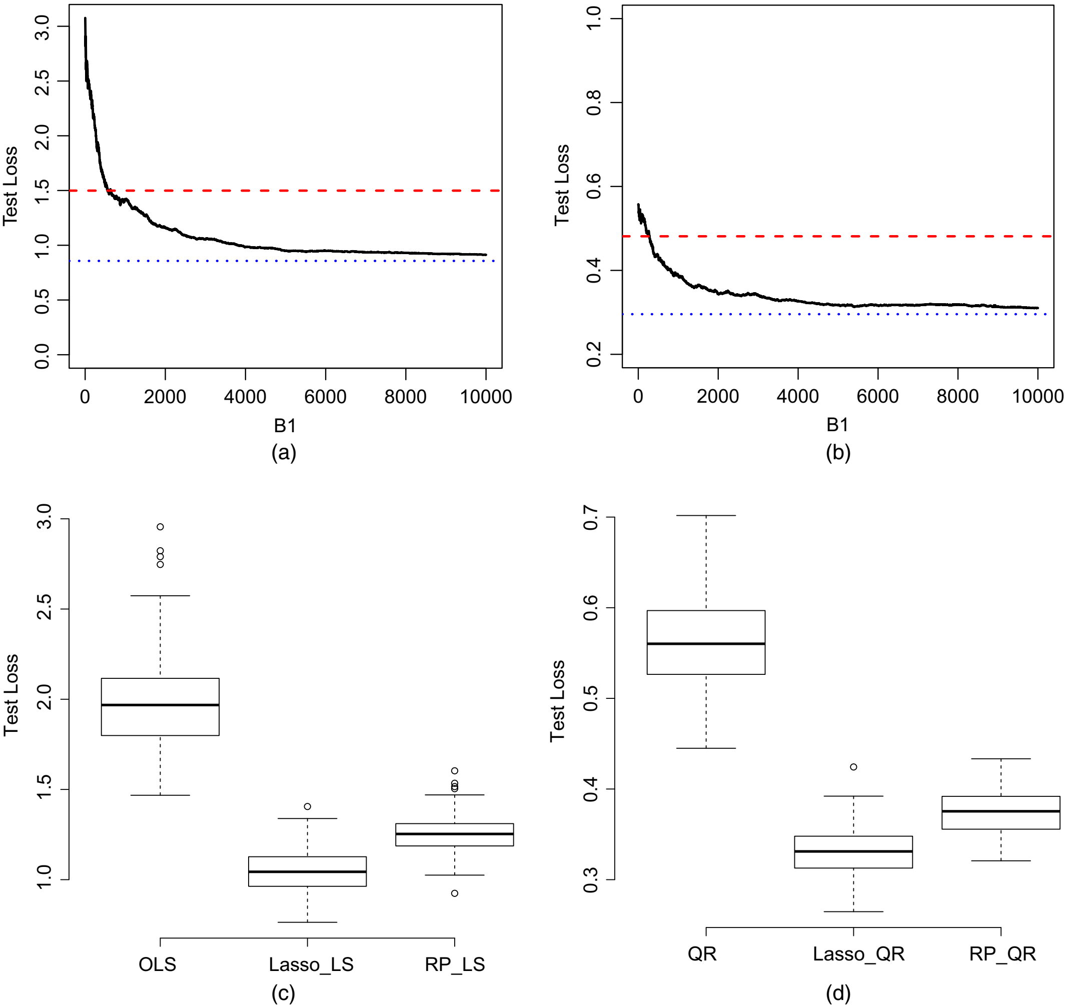


Fig. 9. Test risk as a function of B_1 (with a fixed $B_2 = 20$) for (a) conditional mean regression and (b) conditional quantile regression (– – –, LS; · · · · · ·, LASSO_LS; ———, RP_LS), and averaged test risk over 100 independent repetitions with $B_1 = 2000$ and $B_2 = 20$ for (c) conditional mean regression and (d) conditional quantile regression

for B_1 different random projections, $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$. Here $\mathbf{A}_{b_1}$ is chosen to be the best performer out of B_2 independent projections for each $b_1 = 1, 2, \ldots, B_1$. Now, the RPE regression estimator is

$$f_n^{\mathrm{RP}}(\mathbf{x}) = \frac{\nu_n(\mathbf{x}) - \alpha_m}{\alpha_s},$$

where the constants $\alpha_m \in \mathbb{R}$ and $\alpha_s \in \mathbb{R}^+$ play a similar role to that of α, the threshold value in the RPE classifier. We choose $(\hat{\alpha}_m, \hat{\alpha}_s) = \arg\min_{\alpha_m, \alpha_s} R_n(f_n^{\mathrm{RP}})$.

We consider the model $y_i = \beta^{\mathrm{T}} \mathbf{x}_i + \epsilon_i$, $i = 1, \ldots, n$, where $\beta = (1, 1, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$ and components of $\mathbf{x}_i$ and ϵ_i are independently generated from $N(0, 1)$. We set $p = 100$ and use training and test sets of sizes 200 and 800. We consider both squared loss for mean regression, LS, and check loss for quantile regression, QR, to estimate the conditional 75th percentile. Figs 9(a) and 9(b) show empirical test risks of LS and QR respectively as a function of B_1 (with $B_2 = 20$). Figs 9(c) and 9(d) show averaged test risks for LS and QR respectively over 100 repetitions with $B_1 = 2000$ and $B_2 = 20$. Both are promising for sufficiently large B_1. The true β is sparse and it favours the lasso.

Lastly, we would like to point out a connection to Halko *et al.* (2011) since both methods begin with something random but end up identifying something meaningful.

Julian Stander and Luciana Dalla Valle (*University of Plymouth*)
We congratulate Cannings and Samworth on their paper and R package that makes the random-projection ensemble (RPE) methodology readily applicable. Here we outline an experiment and ask two questions.

We worked with data discussed by Baldino (2016) comprising 120 trip advisor reviews and each reviewer's star classification. We combined one, two and three stars into class 1, and four and five stars into class 2. Using R's `tm` package (Feinerer and Hornik, 2015), we computed the transpose of the term document matrix. This word count matrix had $n=120$ rows corresponding to reviews and $p=2644$ columns corresponding to words, with 97% 0s. We normalized the rows by dividing by review lengths. We randomly selected 60 reviews as training data, with the remaining 60 being test data. We applied the RPE method to the normalized word count feature matrix. For comparison, using dictionaries of 2006 positive and 4783 negative words (Liu *et al.*, 2005), we calculated a sentiment score for each review as the difference between the number of positive and negative word matches. We normalized these scores by dividing by review length to obtain sentiment intensities. We then applied binary logistic regression with sentiment intensity as explanatory variable. Over 50 repetitions, with $d=2$ we obtained quite a low average misclassification rate of 25.5% (standard deviation 1.2%) for the normalized word count RPE methodology using $B_1=500$, $B_2=50$, linear discriminant analysis, Gaussian projection and the leave-one-out test error. The average $\hat{\alpha}$ was 1.69 (0.0127). When quadratic discriminant analysis or the axis projection method was used, the RPE average misclassification rate was often considerably worse (non-overlapping confidence intervals), although the RPE method seemed quite robust to other choices including the value of $d=3,\ldots,9$. For our sentiment intensity logistic regression the average misclassification rate was lower at 12.2% (0.76%). We therefore conclude that the RPE method can be successfully used to classify hotels by using only review word counts. Naturally, better classification results can be obtained by performing a sentiment analysis which makes use of information from positive and negative word dictionaries.

Can the proportion of the classifications $C^1(x),\ldots,C^{B_1}(x)$ in each category be used to quantify classification uncertainty?

The copula construction (Sklar, 1957) provides flexible multivariate models, with vine copulas (Aas *et al.*, 2009) being used when p is large. Could the use of a classifier defined by using copula densities estimated on low dimensional projections of the original data, perhaps with robust marginal modelling, be a way of exploiting copula flexibility, while avoiding high dimension estimation problems?

Milan Stehlík (*Johannes Kepler University in Linz and University of Valparaíso*) **and Luboš Střelec** (*Mendel University in Brno*)
We congratulate Cannings and Samworth, introducing readers to a challenging world of high dimensional classification.

Here we point out underlying algebraic and topological issues, which can play a crucial role for cases of high dimensional classification. The main drawback of the methodology developed can be the need for a group structure underlying the Haar measure. However, in a high dimensional data classification problem, we should not be surprised if the algebraic underlying structure is less strong than a group; it can be indeed a semigroup or even a less structured monoid. Unlike a group, its elements need not have inverses. Thus there should be some construction of the test at hand, which can say 'yes, Haar measure fits our data cloud well', before it is automatically applied. Otherwise, the drawback of random projection is that highly unstable different random projections may lead to radically different clustering results. There is a well-known approach to invariant statistical models based on groups and this will need more attention for cases $p \gg n$. Consider James (1954), Obenchein (1971) and Francis *et al.* (2016) for orthogonal, linear and finite reflection groups as a starter. Topology (as a limiting geometry) can bring much insight into the proper classifier; here topological aggregation can shed light (see Stehlík (2016)). Considering the appropriate topological approach also may enable researchers to control the discontinuities in the smoothing lemma and thus it can achieve better bounds.

Several functionals can do a good classification job if we have a proper training sample. For example deviation from normality in groups of training samples can give us a good chance to build up a simple classifier (e.g. functional on p-value of a properly chosen robust test for normality; Stehlík *et al.* (2014)). For example for the mice data set one can distinguish between 0 and 1 with robust normality tests LT, RJB, $\mathrm{MC}_{\mathrm{LR}}$ and TTRT2 by the difference of p-values. In particular $\mathrm{MC}_{\mathrm{LR}}$ and TTRT2 are very robust; see Stehlík *et al.* (2014).

Paul Switzer (*Stanford University*)
I wish simply to point out the following early references to the use of random projections for classification: Switzer (1970) and Wright and Switzer (1971).

Måns Thulin (*Uppsala University*)
I congratulate Cannings and Samworth on an impressive work which is sure to have an influence on high dimensional statistics for years to come. My comments concern the choice of d and invariance properties.

I would like to stress that there are examples where the choice of d can greatly affect the performance of the random-projection (RP) classifier. For instance, in the mice example with $n=200$, the errors of the LDA-RP and QDA-RP classifiers when $d=50$ are only 5.5 and 4.3 respectively, which are both substantially lower than the error rates for $d=5$ reported in Table 4. If the cross-validation procedure proposed in Section 5.4 is too costly, in addition to $d=5$, another possible default choice is $d=(n-2)/2$, which has been shown to maximize the power of RP in two-sample tests (Lopes *et al.*, 2012).

The random-subspaces method used by the axis-aligned RP classifier has previously been successfully used for classification using decision trees (Ho, 1998; Breiman, 2001), linear classifiers (Skurichina and Duin, 2002) support vector machines (Tao *et al.*, 2006) and for two-sample testing (Thulin, 2014). A potential advantage of using axis-aligned projections instead of Gaussian or Haar projections is that as long as the base classifier is invariant under linear rescaling of the features the axis-aligned RP classifier will also be invariant under linear rescaling.

However, the fact that the RP classifier is not invariant under linear rescaling when Gaussian or Haar projections are used can perhaps be used to our advantage. Consider the mice example again. For a particular test–training data split, the error rate of the RP-LDA5 classifier is 19.4 when using $n=200$. Multiplying a randomly chosen half of the features by $\frac{1}{10}$ and the other half by 10, while the projections used and the test–training split are kept constant, the misclassification rate of the RP-LDA5 classifier can change to anywhere between 10.8 and 38.9, based on 1000 random selections. As a comparison, if the random projections are changed for the non-scaled data with the same test–training split, the error rate just varies between 17.2 and 22.8, and, if standardized data are used, the error is 17.2. Clearly, some rescalings of the data will yield much more favourable results. A topic for future research may therefore be combining RPs with random rescaling.

Jabed H. Tomal (*University of Toronto Scarborough, Toronto*) **and William J. Welch and Ruben H. Zamar** (*University of British Columbia, Vancouver*)
We congratulate Cannings and Samworth for drawing attention to classification in high dimensional settings and their novel approach based on random subsets of features.

A similar method, *ensemble of phalanxes*, EPX, was developed by Tomal *et al.* (2015), where the features are hierarchically clustered into subsets. Base models are trained by using the respective subsets and ensembled. The selection of the *number* and *size* of the models is automatic and data driven.

For example, consider the cardiac arrhythmia diagnoses data from Section 6.2.6 with 194 variables, 452 observations and two classes of sizes 245 and 207. EPX using random forests (RFs) as base learner and the out-of-bag misclassification error (ME) rate as optimization criterion gives a single model with 191 variables. In this instance EPX data adaptively do not choose an ensemble of subsets but perform some feature selection. Following the assessment methodology in the paper, 200 observations are chosen at random for training and 252 are kept for testing. Repeating 100 times, EPX achieves an average test ME of 22.85 (standard error se $=0.24$). In contrast, the top performing random projection has an ME of 26.31 (se $=0.28$). EPX suggests that the 191 variables belong together in one subset or model, and ensembling is not supported.

The examples in the paper exhibit a balanced number of observations in the classes. Many applications, however, have one much sparser class of interest (fraud detection, market research, drug discovery, Internet search, etc.). With unbalanced classes the ME can be a misleading metric. Instead, such applications are often treated as ranking problems: cases are ranked by their probability of belonging to the sparse class. Measures of the quality of ranking include the average hit rate (AHR) (Wang, 2005) and initial enhancement (IE) (Kearsley *et al.*, 1996). Both criteria are larger the better. Looking at the cardiac arrhythmia data we note that they have one normal class of size 245 and 13 *abnormal classes* with smaller frequencies (Table 16). For illustration, we consider the unbalanced data formed by keeping the 245 *normal* instances and 44 instances with *coronary heart disease*. EPX using RFs as the base classifier and the AHR as the optimization criterion uncovers 11 subsets of the original variables. 100 repeats of tenfold cross-validation to evaluate the ranking gives the results in Table 17. EPX is compared with RFs, a method shown by the authors to be a top performer for this application. EPX wins in terms of the AHR and IE.

Table 16. Cardiac arrhythmia classes and their frequencies

Class	1	2	3	4	5	6	7	8	9	10	14	15	16
Frequency	245	44	15	15	13	25	3	2	9	50	4	5	22

Table 17. Means of APR and IE from 100 random repeats of the fitting process

Metric	*Result for EPX*		*Result for RFs*	
	Mean	*Standard error*	*Mean*	*Standard error*
AHR	0.907	0.0006	0.777	0.0013
IE (shortlist=50)	5.166	0.0073	4.778	0.0091
IE (shortlist=100)	2.881	0.0022	2.744	0.0053

Howell Tong (*University of Electronic Science and Technology of China, Chengdu*)
I welcome this paper as it interacts with machine learning. If I were someone looking at the paper from outside the statistical community, I would say that it presents quite an interesting extension to the body of work on random projections, but the algorithm and experimental sections have room for improvement.

For the algorithm section, a fairly natural comparison would be with a popular subset of neural networks. Consider a multilayered neural network where the first layer has linear activations. This network can be viewed as being initialized with a random projection to dimension d (where d is the number of neurons in the first layer) or, equivalently, N random projections to dimension d/N. The subsequent non-linear layers of the network can be viewed as the base classifier of this paper. Neural networks with word embeddings are an example of such networks: `http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/`. References are contained therein.

These types of neural network are commonly used for high dimensional problems. During training, the standard back-propagation neural network training algorithms can be viewed as simultaneously adjusting the initial random projections and base classifier. Such an approach appears to be more elegant than the piecemeal approach adopted in this paper. It would be helpful to understand the trade-offs between these two methods.

For the experiments the 'real world data sets' are small (only 1000 training examples) and, in most part, fairly low dimensional. Consequently, they are not entirely convincing from a modern machine learning perspective. Even 15 years ago, random-projection papers in machine learning were already using larger data sets than most of those reported in this paper. Therefore, the challenge is a comparison on something of a more realistic size for modern use such as the large movie review data set (25 000 reviews, millions of dimensions; see `http://ai.stanford.edu/∼amaas/data/sentiment/`), the entire modified National Institute of Standards and Technology data set (60 000 examples and 784 dimensions, all highly correlated; see `http://yann.lecun.com/exdb/mnist/`), or similar. Adding neural networks as a competitor method would also be useful. In the contemporary machine learning literature using random projections, data sets of the size 2 million data instances and 50 million dimensions have been used. See, for example, `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.5246&rep=rep1&type=pdf`.

Xin Tong (*University of Southern California, Los Angeles*) **and Jingyi Jessica Li** (*University of California, Los Angeles*)
We congratulate Cannings and Samworth for their innovative and thought-provoking paper. In recent years, many classification methods have been developed for high dimensional settings, where the feature dimension p is comparable with or larger than the sample size n. In the literature, most of the existing work has aimed to build a specific procedure (e.g. screening and penalization approaches) to reduce

model complexity effectively. From a different and novel perspective, Cannings and Samworth's work has developed a theory-backed ensemble classification procedure, which first projects features into many lower dimensional spaces so that 'base' classifiers (e.g. linear discriminant analysis and quadratic discriminant analysis) can be applied to the projected data without any modification, and secondly aggregates these low dimensional classifiers via a proper voting scheme. The paper lays a good foundation that motivates us to think about many questions including the following three.

(a) What is the consequence of relaxing assumption 3? This assumption states: 'There exists a projection $A^* \in \mathcal{A}$ such that

$$P_X[\{x \in \mathbb{R}^p : \eta(x) \geqslant \tfrac{1}{2}\} \Delta \{x \in \mathbb{R}^p : \eta^{A^*}(A^* x) \geqslant \tfrac{1}{2}\}] = 0,$$

where Δ denotes the symmetric difference between sets'. Essentially, this assumption means that there is one projection that leads to an oracle decision boundary essentially the same as the oracle decision boundary in the original feature space. Although this is a reasonable and convenient assumption, it can probably be relaxed, and the discrepancy between the two oracle decision boundaries, i.e. the discrepancy between the original Bayes classifier (in $\mathbb{R}^p$) and the best projected Bayes classifier (in $\mathbb{R}^d$), can perhaps show up in the upper bound of the excess error.

(b) In the paper, the voting threshold α is to mimic α^* in equation (12):

$$\alpha^* = \underset{\alpha' \in [0,1]}{\arg\min} [\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}].$$

This is a very natural choice when the classification target is to minimize the classification error (i.e. risk) and when the empirical proportions $\hat{\pi}_1$ and $\hat{\pi}_2$ are good estimates of π_1 and π_2 respectively. How would the authors choose α when we are interested in a type I–II error weighting different from that implied by the class priors, or when good estimates of π_1 and π_2 are lacking?

(c) This comment is related to the first. If we relax assumption 3 to achieve the best performance bounds in Section 4, we no longer prefer d as small as possible (while validating assumption 3). We expect that the best choice of d will depend on the discrepancy between the two Bayes classifiers in R^p and R^d.

Xiangyu Wang (*Google, Mountain View*) **and Chenlei Leng** (*University of Warwick, Coventry*)

We congratulate Cannings and Samworth for a thought-provoking paper on high dimensional ensemble learning. The authors proposed an ensemble method based on random projection, by collecting in some sense good projections, and showed how it could be applied to linear discriminant analysis, quadratic discriminant analysis and other base classifiers. The theory focused on justifying the theoretical properties of the ensemble learner and quantifying the generalizable error between the infinite simulation version and the Bayes risk under three assumptions. We comment on assumption 2 and assumption 3.

The Johnson–Lindenstrauss lemma enables the dimensionality of the data to be reduced from p to a lower number which is independent of p while preserving the pairwise distances of the data. We are not sure, however, whether assumption 2 implicitly makes the dimension of the projected space depend on p. For models with s sparse features ($s \ll p$), a 'good' projection that reduces the dimension from p to d ($d \geqslant O\{\log(n)\} \geqslant s$) with a close-to-optimal empirical loss needs to concentrate on the sparse s-dimensional space. The probability of sampling such a projection under Haar measure would vanish to 0 quickly when $p \to \infty$ for fixed d. Thus, assumption 2 might fail to hold or we might need a weaker version of assumption 2 to make β depend on p.

Assumption 3 controls the difference between the optimal loss of the ensemble classifier and the Bayes risk. This condition alone does not exclude the possibility that selecting a single optimal classifier could perform better than the ensemble classifier, at least in theory. Thus, it would be interesting to see theoretical results showing advantages of using the ensemble method over the single optimal classifier, perhaps in having smaller sampling variance.

Yannis G. Yatracos (*Cyprus University of Technology, Limassol*)

I congratulate Cannings and Samworth on a stimulating and well-written paper. I shall focus on dimension reduction and the use of pseudovalues which introduce additional randomness in the statistical experiment.

Dimension reduction may cause an increase in misclassification proportion. Let $(X_1, \ldots, X_p)$ be a random vector from mixture density $\gamma f^{(p)} + (1-\gamma) h^{(p)}$, with $S_{f^{(p)}}$ and $S_{h^{(p)}}$ the supports of the densities $f^{(p)}$ and $h^{(p)}$, $0 < \gamma < 1$. When $S_{f^{(p)}} \cap S_{h^{(p)}} = \emptyset$, observations from the mixture are naturally separated into two groups, from $f^{(p)}$ and $h^{(p)}$. When either the vector's components are independently and identically distributed or with mild assumptions on the conditional densities of X_i gives $X_{i-1},\ldots,X_1, 1\leqslant i\leqslant p$,

$$\int_{S_{f^{(p)}}\cap S_{h^{(p)}}\cap\{f^{(p)}\geqslant h^{(p)}\}} h^{(p)}(x)\,\mathrm{d}x+\int_{S_{f^{(p)}}\cap S_{h^{(p)}}\cap\{h^{(p)}\geqslant f^{(p)}\}} f^{(p)}(x)\,\mathrm{d}x\leqslant\rho(f^{(p)},h^{(p)})\downarrow 0,$$

as $p\rightarrow\infty$; $\rho(f^{(p)},h^{(p)})$ is the integral of $\surd(f^{(p)}h^{(p)})$ over $S_{f^{(p)}}\cap S_{h^{(p)}}$. Simulations confirm that misclassification proportions in $f^{(p)}$ and $h^{(p)}$ decrease to 0 as p increases (Yatracos, 2013, 2017). In this respect, large p is a blessing in classification and cluster detection problems.

Random projections are similar in spirit to bootstrap samples. Both introduce additional randomness in the experiment because, in practice, the numbers of random projections and of the bootstrap samples are finite. This causes an additional positive term in the mean-square errors of estimates. Let C^{RP}_{n,B_1} be the *random-projection ensemble* classifier, $R(C^{\mathrm{RP}}_{n,B_1})$ the estimate of $R(C^{\mathrm{Bayes}})$ and $\mathcal{T}_n$ the data, and E denotes expected value:

$$\tilde{R}_{n,B_1}=E[R(C^{\mathrm{RP}}_{n,B_1})|\mathcal{T}_n],\qquad E[\tilde{R}_{n,B_1}]=E[R(C^{\mathrm{RP}}_{n,B_1})].$$

Then the mean-square error of $R(C^{\mathrm{RP}}_{n,B_1})$ is

$$E[R(C^{\mathrm{RP}}_{n,B_1})-R(C^{\mathrm{Bayes}})]^2=E[\tilde{R}_{n,B_1}-R(C^{\mathrm{Bayes}})]^2+E[\mathrm{var}\{R(C^{\mathrm{RP}}_{n,B_1})|\mathcal{T}_n\}].\qquad(51)$$

In equation (51), the *cushion error* $\mathrm{var}\{R(C^{\mathrm{RP}}_{n,B_1})|\mathcal{T}_n\}$ is positive when B_1 is finite. This term vanishes with a jackknife-type approach when a finite number of pseudovalues are available and used, for example, either obtained from *all* 'leave-one-out data subsets' instead of B ($<\infty$) bootstrap samples (Yatracos, 2002), or by considering orthogonal projections in hyperplanes determined by the data vectors $\mathcal{T}_n$, which carry all the information. The latter is useful with 'ultrahigh dimensional settings' (Section 7). For high dimensional normal and t-mixtures, data projections on each observation vector reduced the misclassification proportion and the computational time achieved with ϵ-net projection vectors (Yatracos (2013), page 41).

The **authors** replied later, in writing, as follows.

We are very grateful to the discussants for their insightful comments on our work, and we are glad to find a broad consensus that methods based on random projections offer considerable promise for high dimensional data analysis. The comments are extremely wide ranging, and we apologize in advance for the fact that, for brevity, we cannot address all of them. It is clear, however, that there is considerable scope for future research in this area, and we look forward to witnessing and contributing to its development.

Correlation between features

Kent presents an interesting toy example, which focuses on the effect of the correlation between the features. As we discuss in Fig. 1 of the main text, it is usually only sensible to aggregate over carefully selected (rather than all) projections. Even in Kent's high correlation case ($\rho=0.99$), where only 5% of projections result in a base classifier with at least half the discriminatory power, we still expect with $B_2=50$ to find such a projection in most groups. We carried out a small simulation study on Gaussian class conditional distributions with $\pi_0=\pi_1=\frac{1}{2}$:

(a) case 1a, $p=2$, $\rho=0$, $\mu_1=a_1(1,0)^{\mathrm{T}}$ and $\mu_0=a_1(-1,0)^{\mathrm{T}}$, where a_1 is such that the Bayes risk is 14.44%;
(b) case 1b, $p=2$, $\rho=0.99$, $\mu_1=a_2(1,-1)^{\mathrm{T}}$ and $\mu_0=a_2(-1,1)^{\mathrm{T}}$, where a_2 is such that the Bayes risk is 14.44%.

In Table 18 we present the misclassification errors for linear discriminant analysis (LDA) applied to the original data and the random-projection ensemble classifier with $d=1$, $B_1=500$, $B_2=50$ and $n=200$, and both Gaussian and axis-aligned projections. We also present the average test error of the LDA classifier applied on the chosen projections. LDA is tailored to these set-ups, and indeed it performs very well; the RP-LDA$_1$ classifier has similar performance in both cases. The extreme correlation ($\rho=0.99$) does not greatly affect the performance of the RP-LDA$_1$ (Gaussian) classifier; in particular, although the high correlation does have a small effect on the average error base classifier applied on the chosen projections, this is overcome in the ensemble step. This illustrates what we believe to be the advantage of aggregation over the choice of a single projection (discussed by de Carvalho, Page and Barney).

Table 18. Misclassification rates for the Gaussian toy example

Case	*LDA*	*Results for Gaussian projection*		*Results for axis-aligned projection*	
		RP-LDA$_d$	$B_1^{-1}\sum_{b_1=1}^{B_1} R(C_n^{\mathbf{A}_{b_1}})$	*RP-LDA*$_d$	$B_1^{-1}\sum_{b_1=1}^{B_1} R(C_n^{\mathbf{A}_{b_1}})$
1a	$14.2_{0.2}$	$15.1_{0.4}$	$15.3_{0.5}$	$14.2_{0.3}$	$14.2_{0.3}$
1b	$14.8_{0.3}$	$15.3_{0.3}$	$17.6_{0.3}$	$47.1_{0.4}$	$47.1_{0.4}$
2a	$27.1_{0.8}$	$19.7_{0.6}$	$38.4_{0.3}$	$14.9_{0.6}$	$18.0_{0.4}$
2b	$27.7_{0.9}$	$21.6_{0.9}$	$38.8_{0.3}$	$19.4_{0.8}$	$25.1_{0.3}$

We now repeat the experiment with $p=100$ and $d=5$, and all other parameters kept as before. The class conditional covariance matrices have 1s on the diagonal and ρ on the off-diagonal:

(a) case 2a, $p=100, \rho=0, \mu_1=a_3(1,0,\ldots,0)^{\mathrm{T}}$ and $\mu_0=a_3(-1,0,\ldots,0)^{\mathrm{T}}$, where a_3 is such that the Bayes risk is 14.44%;

(b) case 2b, $p=100, \rho=0.99, \mu_1=a_4(1,-1,0,\ldots,0)^{\mathrm{T}}$ and $\mu_0=a_4(-1,1,0,\ldots,0)^{\mathrm{T}}$, where a_4 is such that the Bayes risk is 14.44%.

Here, the sample covariance matrix is ill conditioned, so LDA performs poorly, and the random-projection ensemble classifier offers considerable improvement. In cases 1a, 2a and 2b, assumption 3 holds with an axis-aligned projection. The axis-aligned version performs better here since we restrict the set of projections, so we have a greater chance of finding good ones. However, in case 1b there is no axis-aligned projection that results in a classifier that is significantly better than a random guess, and the resulting random-projection ensemble classifier is also close to a random guess.

Methodological variations

Many discussants suggested alternatives to our basic methodological proposal. These included the assignment of weights to the selected projections, based on their empirical performance (Chen and Shah, Feng, Zhang, and Josh, Fan and James), choosing projections via projection pursuit (Janson), consideration of the underlying algebraic and topological structure (Stehlík and Střelec), decoupling rotation and dimension reduction (Blaser and Fryzlewicz) or averaging over class probability estimates rather than classifiers (Gneiting and Lerch). These are attractive and sensible ideas, though, similarly to Chen and Shah, we found in our experiments that more sophisticated weighting schemes led to only relatively minor (if any) improvements. One advantage of our proposal is that it can be analysed theoretically, through the independence of the selected projections, conditional on the training data. Meanwhile, Tomal, Welch and Zamar highlight their ensemble-of-phalanxes method, where features are clustered hierarchically into subsets, Casarin, Frattorolo and Rossini, and Stander and Dalla Valle suggest copula-based discriminant analysis and Tong discusses neural network approaches, which are also attractive but currently seem less amenable to theoretical understanding.

Some contributors discussed the axis-aligned version of our proposal in more detail (Janson, and Ling, Yang and Xue). Another popular alternative was to generate the projections from different distributions with the aim of finding *good* projections more efficiently (Blaser and Fryzlewicz, Zhang, and Derenski, Fan and James). Other ideas included choosing new projections to be dissimilar to those already chosen; either orthogonal (Feng) or by adding some similarity penalty (Lu and Xue). We remark that, in our experience and in high dimensions, the selected projections tend to be nearly orthogonal anyway. Thulin suggests including a random rescaling when generating the projections; in contrast, both Critchley and Durrant discuss deterministic rescaling or standardizing of the variables. Although one could construct examples where such renormalization would lead to poor performance, these ideas are certainly worth investigating further.

Our paper focuses on 0–1 error loss, where the two types of misclassification are assumed equally serious. As pointed out by both Hand, and Tong and Li, in practice often one type of error is more serious than the other. Suppose now that, for some $m>0$,

$$R(C)=\pi_1\int_{\mathbb{R}^p}\mathbb{1}_{\{C(x)=0\}}\,\mathrm{d}P_1(x)+m\pi_0\int_{\mathbb{R}^p}\mathbb{1}_{\{C(x)=1\}}\,\mathrm{d}P_0(x),$$

so that assigning a class 0 observation to class 1 is m times more serious than the other error. Three modifications should be made to the methodology. First, the base classifier should target the misclassification imbalance; for example, for LDA the projected data base classifier would be

$$C_n^{A-\mathrm{LDA}}(x) := \begin{cases} 1 & \text{if } \log\left(\dfrac{\hat{\pi}_1}{m\hat{\pi}_0}\right) + \left(Ax - \dfrac{\hat{\mu}_1^A + \hat{\mu}_0^A}{2}\right)^{\mathrm{T}} \hat{\Omega}^A(\hat{\mu}_1^A - \hat{\mu}_0^A) \geqslant 0. \\ 0 & \text{otherwise.} \end{cases}$$

Second, the projections should be selected on the basis of the corresponding weighted estimate (see equation (7) in the main text), for example using the training error

$$R_n^A := \frac{1}{n_1 + mn_0}\left[\sum_{\{iP:Y_i=1\}} \mathbb{1}_{\{C_n^{\Lambda}(X_i)=0\}} + m \sum_{\{i:Y_i=0\}} \mathbb{1}\{C_n^{\Lambda}(X_i)=1\}\right].$$

Finally, α should be chosen to mimic the weighted version of equation (5), i.e.

$$\alpha^* = \underset{\alpha' \in [0,1]}{\arg\min}[\pi_1 G_{n,1}(\alpha') + m\pi_0\{1 - G_{n,0}(\alpha')\}].$$

Theoretical extensions

Several discussants (Critchley, Fan and Zhu, Feng, Kong, Shi, Song and Lu, Tong and Li, and Wang and Leng) comment on our theoretical assumptions, and in particular the quantity β in our assumption 2. Since the training data are considered fixed in the corresponding section of the paper, β can depend on the training data (and therefore n and p). In the on-line supplement, we show that in practice we can typically expect assumption 2 to hold with β not too small. We see in particular that increasing p does not necessarily lead to $\beta \searrow 0$ (recall that the Johnson–Lindenstrauss lemma guarantees that, regardless of the magnitude of p, we can reduce dimension from p to $O\{\log(n)\}$ while nearly preserving pairwise distances).

Assumption 3 is at the population level. A natural relaxation is to assume that the oracle projection A^* does not perfectly preserve the class information, but instead to allow for a region where the projected classifier disagrees with the Bayes classifier. This can be formalized through the existence of a projection $A^* \in A$ and $\tau \geqslant 0$ such that

$$P_X(\{x \in \mathbb{R}^p : \eta(x) \geqslant \tfrac{1}{2}\} \Delta \{x \in \mathbb{R}^p : \eta^{A^*}(A^*x) \geqslant \tfrac{1}{2}\}) = \tau.$$

Then, by a straightforward extension to proposition 2, we have that $R(C^{\mathrm{Bayes}}) \leqslant R(C^{A^*-\mathrm{Bayes}}) \leqslant R(C^{\mathrm{Bayes}}) + \tau$.

Bing and Wegkamp suggest a possible alternative approach to our theoretical analysis, which involves regarding the random-projection classifier as a *plug-in* rule with $\nu_n(x) + \frac{1}{2} - \alpha$ acting as an estimate of $\eta(x)$. We have found that $\nu_n(x)$ is not a good estimate of $\eta(x)$ (even with the suggested bias correction), though it would be interesting to find conditions under which we can hope to estimate η by using our random-projection methodology (see Gneiting and Lerch).

Numerical comparisons

We welcome the contributions which added to our numerical work, aiding the understanding of the practical properties of the random-projection ensemble classifier. For instance, Gallaugher and McNicholas compare with mixture discriminant analysis, whereas Stander and Dalla Valle apply the random-projection ensemble classifier to a trip advisor data set.

Hennig and Viroli found that our proposal performed poorly compared with their quantile-based classifier in two of their set-ups. In their set-up 2, class 1 has p independent, log-normal components, whereas (in the $100q\%$ signal variables case) class 0 has p independent components, qp log-normal components shifted by 0.2 and $(1-q)p$ log-normal components. A key characteristic of the data in this set-up is that all variables are skewed and positive. In this example, our assumption 3 does not hold for $d=5$, and in fact the best low dimensional projection has high test error (compared with a Bayes risk of almost 0 when $q=1$). Nevertheless, we can check for skewness and include a marginal logarithmic transformation as a preprocessing step in this instance. In Table 19, we present error rates when data are generated from Hennig and Viroli's set-up 2 with $p=100$ and $n=50$, and we take componentwise logarithms of the data before applying the random-projection methodology. For reference we also present the performance of the quantile-based methods QCG and QCS from Hennig and Viroli's discussion. Our transformation works very well when $q=1$ (it should be noted that many of the other methods discussed by Hennig and Viroli

Table 19. Misclassification rates for the random-projection ensemble classifier for set-up 2 with log-preprocessing ($B_1=500$; $B_2=50$; Gaussian projections)

q	*Misclassification rates for the following classifiers:*				
	RP-LDA$_5$	*RP-QDA*$_5$	*RP-knn*$_5$	*QCG*	*QCS*
1	$18.4_{0.9}$	$12.6_{1.6}$	$16.4_{2.2}$	25.7	21.3
0.1	$46.7_{0.7}$	$46.4_{0.6}$	$46.1_{0.5}$	44.3	41.5

may also benefit from this preprocessing). In the case $q=0.1$ and when n is this small, the problem is very challenging and all methods struggle; in particular, we are unable to retain many of the signal projections because our overfitting term ϵ_n is large.

Bergsma and Jamil use only $B_1=30$ and $B_2=5$ when using the random-projection methodology in conjunction with Gaussian process regression with fractional Brownian motion for reasons of computational cost. We have found that larger values of B_1 and B_2 give considerably better results, but fortunately simple (and quick-to-compute) base classifiers usually suffice. Hand suggests a comparison with a weighted k-nearest-neighbour classifier. One option is the bagged nearest neighbour classifier, which is essentially a weighted nearest neighbour classifier with geometrically decaying weights (Hall and Samworth, 2005; Biau and Devroye, 2010). An alternative is to use the optimal weighting scheme, which produces an asymptotic improvement of 5–10% in excess risk over the unweighted k-nearest-neighbour classifier when $d\leqslant 15$ (Samworth, 2012). It would be interesting to see whether similar improvements are obtained when used in conjunction with the random-projection methodology.

Other statistical problems

It was particularly pleasing to see many contributions that discuss using the random-projection ensemble framework to tackle other high dimensional statistical problems. Several contributors suggested ways in which the information in the chosen projections can be aggregated to provide measures of variable importance (Anderlucci, Montanari and Fortunato, Derenski, Fan and James, and Gataric). Li and Yu, Critchley, and Murtagh and Contreas considered clustering (unsupervised learning) problems, where the labels of the training data are unknown. Here we require both a (sample) measure of the performance of the base method to select the projections analogously to equation (7) in the main text, and a suitable method for aggregating the chosen projections. Fan and Zhu discuss the use of random projections for the estimation of the top k left singular space of a data matrix; the result they state together with an appropriate version of Wedin's theorem (Wedin, 1972; Yu *et al.*, 2015; Wang, 2016) may allow the control of the sine angle distance they seek. Other interesting new directions discussed include interaction network learning (Demirkaya and Lv), regression (Kong, and Shin, Zheng and Wu), feature detection (Mateu) and estimation of central subspaces in the context of sufficient dimension reduction (Sabolová and Marriott).

Which random-projection ensemble classifier?

We are greatful to Switzer for pointing out two early references to the use of random projections for classification. As noted by some discussants (Hand, Hennig and Viroli, and Critchley), the flexibility offered by our random-projection ensemble classification framework naturally poses the question of when a particular base classifier should be used (of course, analogous questions arise regardless of whether methods are used in conjunction with random projections). If no natural choice is suggested from understanding of the data-generating mechanism, one possible approach is to randomize the choice of base classifier for each projection, say choosing between LDA, quadratic discriminant analysis and k-nearest neighbours, each with probability $\frac{1}{3}$. Alternatively, we can try all three base methods on each projection and retain the projection, base method pair that minimizes the leave-one-out error estimate. If one of these three original classifiers is clearly best, then it should emerge as the 'winner' within most groups of B_2 projections. This strategy therefore provides additional robustness, and the theory goes through unchanged for these versions of the random-projection ensemble classifiers. Post-pruning, as suggested by Fortunato, is another option, but we do not pursue that here. We implement both methods proposed above (denoted RP-Random$_d$ and RP-Min$_d$) in a small simulation study, summarized in Table 20, where the model numbers refer to the settings described in Section 6.1. For models 2, 3 and 4, the risks of both variants of

Table 20. Misclassification rates for the randomized and selected base classifier variants, with $p = 100$, $n = 200$, $B_1 = 500$, $B_2 = 50$, $d = 5$ and Gaussian projections†

Classifier	*Misclassification rates for the following classifiers:*			
	Model 1	*Model 2*	*Model 3*	*Model 4*
RP-Random$_5$	$26.2_{0.7}$	$6.0_{0.2}$	$3.6_{0.1}$	$23.6_{0.5}$
RP-Min$_5$	$23.6_{0.7}$	$6.1_{0.3}$	$3.7_{0.2}$	$23.9_{0.6}$
Best RP	$22.32_{0.32}$	$5.58_{0.12}$	$4.23_{0.14}$	$24.02_{0.30}$

†For comparison, in the bottom row we present the risk of the best performing version of the random-projection ensemble classifier as seen in Section 6.1 of the main text.

the classifier are comparable with (or better than) that of the best performing choice of base method. For model 1 there is only a slight deterioration in performance. Taking these ideas further, and addressing comments from Bing and Wegkamp, Critchley, and Liu and Cheng, one could even add randomization over d and/or Gaussian or axis-aligned projections.

Ultrahigh dimensional problems

Tong discusses the applicability of our random-projection methodology in contemporary machine learning problems. He correctly points out that some modern data sets have potentially millions of features and observations, far larger than the problem sizes we investigate in our numerical studies in Section 6. Of course, the fact that such large data sets exist does not mean that we should neglect the (still relevant) smaller problems. Moreover, in ultrahigh dimensional problems it is often reasonable to assume that only a subset of the features are relevant. Indeed, many studies of such problems focus on reducing the data dimension by attempting to screen out the noise variables (e.g. Fan and Lv (2008), Fan *et al.* (2009), Meinshuasen and Bühlmann (2010) and Shah and Samworth (2013)). If high dimension is still a problem, another common technique is to use a single random projection (e.g. Achlioptas (2003)) into a lower dimensional space. Either or both of these techniques can be used as a preprocessing step to give thousands, say, rather than millions of features, and then the random-projection methodology can be applied. In fact, in the paper by Dahl *et al.* (2013) cited by Tong, to make the problem more manageable, the authors apply feature screening and a sparse random projection to reduce the dimension to 4000, before applying a neural net classifier.

Responses to direct questions

Gallaugher and McNicholas seek clarification about our real data settings—we used the hill–valley data set without noise, pooled the training and test sets, and then subsampled at random our own training and test sets as described in Section 6.2. The missing values in the mice data set were imputed as the sample average value for that feature for the non-missing entries. Kong asks why the performance improves as p increases for model 1. One reason is that, although the signal is the same (the Bayes risk is 4.45% in both cases), the variance of the noise components is reduced in the higher dimensional setting; see also the explanation of Yatracos. In answer to Zhang, penalized logistic regression does not perform well in setting 1 because, despite the fact that the model is highly sparse (only two features are relevant for classification), the class boundaries are non-linear. Stander and Dalla Valle ask whether it is possible to quantify classification uncertainty by using $C_n^{A_1}, \ldots, C_n^{A_{B_1}}$. Regarding the training data as fixed and having observed $\nu_n(x) = t < \alpha$, say, one can indeed obtain a simple bound on the probability of observing $\nu_n(x)$ at least as small as t when $C_n^{\mathrm{RP}*}(x) = 1$ (a kind of 'p-value'), via the fact that $\nu_n(x) \sim B_1^{-1}\mathrm{Bin}\{B_1, \mu_n(x)\}$.

References in the discussion

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009) Pair-copula constructions of multiple dependence. *Insur. Math. Econ.*, **44**, 182–198.

Achlioptas, D. (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comp. Syst. Sci.*, **66**, 671–687.

Altham, P. M. E. (1978) Two generalizations of the binomial distribution. *Appl. Statist.*, **27**, 162–167.

Altman, E. (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finan.*, **23**, 589–609.

Baldino, A. (2016) Information mining from social media. *Master's Thesis*. University of Rome La Sapienza, Rome.

Bassetti, F., Casarin, R. and Ravazzolo, F. (2017) Bayesian nonparametric calibration and combination of predictive distributions. *J. Am. Statist. Ass.*, to be published.

Benton, T. (2002) Theoretical and empirical models. *PhD Thesis*. Department of Mathematics, Imperial College London, London.

Biau, G. and Devroye, L. (2010) On the layered nearest neighbour estimate, the bagged, nearest neighbour estimate and the random forest method in regression and classification. *J. Multiv. Anal.*, **101**, 2499–2518.

Bishop, C. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st edn. New York: Springer.

Blaser, R. and Fryzlewicz, P. (2016) Random rotation ensembles. *J. Mach. Learn. Res.*, **17**, 1–26.

Bohorquez, M., Giraldo, R. and Mateu, J. (2016) Optimal dynamic spatial sampling. *Environmetrics*, **27**, 293–305.

Bohorquez, M., Giraldo, R. and Mateu, J. (2017) Multivariate functional random fields: prediction and optimal sampling. *Stoch. Environ. Res. Risk Assessmnt*, **31**, 53–70.

Breiman, L. (1996) Stacked regressions. *Mach. Learn.*, **24**, 49–64.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Cook, R. D. and Critchley, F. (2000) Identifying regression outliers and mixtures graphically. *J. Am. Statist. Ass.*, **95**, 781–794.

Critchley, F. and Heiser, W. (1988) Hierarchical trees can be perfectly scaled in one dimension. *J. Classificn*, **5**, 5–20.

Dahl, G. E., Stokes, J. W., Deng, L. and Yu, D. (2013) Large-scale malware classification using random projections and neural networks. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, *Vancouver*, pp. 3422–3426.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Dietterich, T. G. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach. Learn.*, **40**, 139–157.

Duin, R. P. W. (1996) A note on comparing classifiers. *Pattn Recogn Lett.*, **17**, 529–536.

Durrant, R. J. and Kabán, A. (2015) Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Mach. Learn.*, **99**, 257–286.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016) Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings (with discussion). *J. R. Statist. Soc.* B, **78**, 505–562.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*, 5th edn. New York: Wiley.

Fan, J. and Fan, Y. (2008) High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015) Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist.*, **43**, 1243–1272.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, Y. and Lv, J. (2016) Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *Ann. Statist.*, **44**, 2098–2126.

Fan, J., Samworth, R. and Wu, Y. (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.

Fang, H.-B., Fang, K.-T. and Kotz, S. (2002) The meta-elliptical distributions with given marginals. *J. Multiv. Anal.*, **82**, 1–16.

Feinerer, I. and Hornik, K. (2015) tm: text mining package. *R Package Version 0.6-2*. (Available from `http://CRAN.R-project.org/package=tm`.)

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Ass.*, **97**, 611–631.

Fraley, C., Raftery, A. E. and Scrucca, L. (2017) mclust: normal mixture modeling for model-based clustering, classification, and density estimation. *R Package Version 5.2.3*.

Francis, A., Stehlík, M. and Wynn, H. (2017) Building exact confidence nets. *Bernoulli*, **23**, 3145–3165.

Friedman, J. H. and Stuetzle, W. (198l) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543–552.

Ghahramani, Z. and Hinton, G. E. (1997) The EM algorithm for factor analyzers. *Technical Report CRG-TR-96-1*. University of Toronto, Toronto.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Guhaniyogi, R. and Dunson, D. B. (2015) Bayesian compressed regression. *J. Am. Statist. Ass.*, **110**, 1500–1514.

Halko, N., Martinsson, P.-G. and Tropp, J. A. (2011) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, **53**, 217–288.

Hall, P. and Samworth, R.J. (2005) Properties of bagged nearest neighbour classifiers. *J. R. Statist. Soc.* B, **67**, 363–379.

Hall, P., Titterington, D. M. and Xue, J. H. (2009) Median-based classifiers for high-dimensional data. *J. Am. Statist. Ass.*, **104**, 1597–1608.

Hall, P. and Xue, J.-H. (2014) On selecting interacting features from high-dimensional data. *Computnl Statist. Data Anal.*, **71**, 694–708.

Han, F., Zhao, T. and Liu, H. (2013) CODA: high dimensional copula discriminant analysis. *J. Mach. Learn. Res.*, **14**, 629–671.

Hand, D. J. (1997) *Construction and Assessment of Classification Rules*. Chichester: Wiley.

Hand, D. J. (2006) Classifier technology and the illusion of progress. *Statist. Sci.*, **21**, 1–14.

Hastie, T. and Tibshirani, R. (1986) Generalized additive models. *Statist. Sci.*, **1**, 297–310.

Hastie, T. and Tibshirani, R. (1996) Discriminant analysis by Gaussian mixtures. *J. R. Statist. Soc.* B, **58**, 155–176.

He, Y., Zhang, X. and Wang, P. (2016) Discriminant analysis on high dimensional Gaussian copula model. *Statist. Probab. Lett.*, **117**, 100–112.

Hennig, C. and Viroli, C. (2016) Quantile-based classifiers. *Biometrika*, **103**, 435–446.

Herbei, R. and Wegkamp, M. H. (2006) Classification with reject option. *Can. J. Statist.*, **34**, 709–721.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. R. (2012) Improving neural networks by preventing co-adaptation of feature detectors. *Preprint arXiv:1207.0580*.

Ho, T. K. (1998) The random subspace method for constructing decision forests. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 832–844.

Hoadley, B. (2001) Comment on “Statistical modelling: the two cultures”. *Statist. Sci.*, **16**, 220–224.

Holte, R. (1993) Very simple classification rules perform well on most commonly used data sets. *Mach. Learn.*, **11**, 63–91.

Huber, P. J. (1985) Projection pursuit (with discussion and rejoinder). *Ann. Statist.*, **13**, 435–525.

Jamain, A. (2004) Meta-analysis of classification methods. *PhD Thesis*. Department of Mathematics, Imperial College London, London.

Jamain, A. and Hand, D. J. (2008) Mining supervised classification performance studies: a meta-analytic investigation. *J. Classificn*, **25**, 87–112.

James, A. T. (1954) Normal multivariate analysis and the orthogonal group. *Ann. Math. Statist.*, **25**, 40–75.

Jiang, B. and Liu, J. S. (2014) Variable selection for general index models via sliced inverse regression. *Ann. Statist.*, **42**, 1751–1786.

Kaski, S. (1998) Dimensionality reduction by random mapping: fast similarity computation for clustering. In *Proc. Int. Jt Conf. Neural Networks*, pp. 413–418.

Ke, Y., Fu, B. and Zhang, W. (2016) Semi-varying coefficient multinomial logistic regression for disease progression risk prediction. *Statist. Med.*, **35**, 4764–4778.

Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T. and Sheridan, R. P. (1996) Chemical similarity using physiochemical property descriptors. *J. Chem. Informn Computnl Sci.*, **36**, 118–127.

Kong, Y., Li, D. and Lv, J. (2016) Interaction pursuit in high-dimensional multi-response regression via distance correlation. *Ann. Statist.*, to be published.

Kuncheva, L. I. and Whitaker, C. J. (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, **51**, 181–207.

Li, B. (2006) A new approach to cluster analysis: the clustering-function-based method. *J. R. Statist. Soc.* B, **68**, 457–476.

Li, D., Ke, Y. and Zhang, W. (2015) Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *Ann. Statist.*, **43**, 2676–2705.

Li, N., Yu, Y. and Zhou, Z.-H. (2012) Diversity regularized ensemble pruning. In *Proc. Jt Eur. Conf. Machine Learning and Knowledge Discovery in Databases*, pp. 330–345. Berlin: Springer.

Liu, B., Hu, M. and Cheng, J. (2005) Opinion observer: analyzing and comparing opinions on the web. In *Proc. 14th Int. Conf. World Wide Web, Chiba, May 10th–14th*, pp. 342–351. New York: Association for Computing Machinery.

Lopes, M. E., Jacob, L. J. and Wainwright, M. J. (2012) A more powerful two-sample test in high dimensions using random projection. *Preprint arXiv:1108.2401v2*.

Marriott, P., Sabolová, R., Van Bever, G. and Critchley, F. (2015) Geometry of goodness-of-fit testing in high dimensional low sample size modelling. In *Geometric Science of Information: Proc. 2nd Int. Conf.* (eds F. Nielsen and F. Barbaresco), pp. 596–576. New York: Springer.

Marriott, P., Sabolová, R., Van Bever, G. and Critchley, F. (2016) The information geometry of sparse goodness-of-fit testing. *Entropy*, **18**, 421–440.

Mateu, J., Lorenzo, G. and Porcu, E. (2007) Detecting features in spatial point processes with clutter via local indicator of spatial association. *J. Computnl Graph. Statist.*, **16**, 968–990.

Mateu, J., Lorenzo, G. and Porcu, E. (2010) Features detection in spatial point processes via multivariate techniques. *Environmetrics*, **21**, 400–414.

McNicholas, P. D. (2016) *Mixture Model-based Classification*. Boca Raton: Chapman and Hall–CRC.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Montanari, A. and Lizzani, L. (2001) A projection pursuit approach to variable selection. *Computnl Statist. Data Anal.*, **35**, 463–473.

Murtagh, F. (2016) Sparse p-adic data coding for computationally efficient and effective Big Data analytics. *p-Adic Numbrs Ultrametr. Anal. Appl.*, **8**, 236–247.

Murtagh, F. (2017) *Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics*. Boca Raton: Chapman and Hall–CRC.

Murtagh, F. and Contreras, P. (2015) Random projection towards the Baire metric for high dimensional clustering. In *Statistical Learning and Data Sciences* (eds A. Gammerman, V. Vovk and H. Papadopoulos), pp. 424–431. New York: Springer.

Murtagh, F. and Contreras, P. (2017) Clustering through high dimensional data scaling: applications and implementations. *Arch. Data Sci.* A, **2**, 1–16.

Obenchein, R. L. (1971) Multivariate procedure invariant under linear transformations. *Ann. Math. Statist.*, **42**, 1569–1578.

Page, G. L., Bhattacharya, A. and Dunson, D. B. (2013) Classification via Bayesian nonparametric learning of affine subspaces. *J. Am. Statist. Ass.*, **108**, 187–201.

Park, M. Y. and Hastie, T. J. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30–50.

Ranjan, R. and Gneiting, T. (2010) Combining probability forecasts. *J. R. Statist. Soc.* B, **72**, 71–91.

Rodriguez, J. J., Kuncheva, L. I. and Alonso, C. J. (2006) Rotation forest: a new classifier ensemble method. *IEEE Trans. Pattn Anal. Mach. Intell.*, **28**, 1619–1630.

Samworth, R. J. (2012) Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, **40**, 2733–2763.

Schclar, A. and Rokach, L. (2009) Random projection ensemble classifiers. In *Enterprise Information Systems: Proc. 11th Int. Conf. Milan, May 6th–10th* (eds J. Filipe and J. Cordeiro), pp. 309–316. Berlin: Springer.

Segers, J., van den Akker, R. and Werker, B. J. M. (2014) Semiparametric Gaussian copula models: geometry and efficient rank-based estimation. *Ann. Statist.*, **42**, 1911–1940.

Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: another look at stability selection. *J. R. Statist. Soc.* B, **75**, 55–80.

Sklar, A. (1957) Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.

Skurichina, M. and Duin, R. P. W. (2002) Bagging, boosting and the random subspace method for linear classifiers. *Pattn Anal. Appl.*, **5**, 121–135.

Stehlík, M. (2016) On convergence of topological aggregation functions. *Fuzzy Sets Syst.*, **287**, 48–56.

Stehlík, M., Thulin, M. and Střelec, L. (2014) On robust testing for normality in chemometrics. *Chemometr. Intell. Lab. Syst.*, **130**, 98–108.

Sun, W. W., Qiao, X. and Cheng, G. (2016) Stabilized nearest neighbor classifier and its statistical properties. *J. Am. Statist. Ass.*, **111**, 1254–1265.

Switzer, P. (1970) Numerical classification. In *Computer Applications in the Earth Sciences: Geostatistics* (ed. D. F. Merriam), pp. 31–43. New York: Springer.

Tao, D., Tang, X., Li, X. and Wu, X. (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattn Anal. Mach. Intell.*, **28**, 1088–1099.

Thulin, M. (2014) A high-dimensional two-sample test for the mean using random subspaces. *Computnl Statist. Data Anal.*, **74**, 26–38.

Tibshirani, R. J., Hastie, T. J., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natn. Acad. Sci. USA*, **99**, 6567–6572.

Tsybakov, A. B. (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, **32**, 135–166.

Wager, S., Wang, S. and Liang, P. S. (2013) Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems* (eds C. J. C Burges, L. Bottou, M. Welling, Z.Ghahramani and K. Q. Weinberger), vol. 26, pp. 351–359. Red Hook: Curran Associates.

Wang, T. (2016) Spectral methods and computational trade-offs in high-dimensional statistical inference. *PhD Thesis*. University of Cambridge, Cambridge.

Wang, Q., Yin, X. and Critchley, F. (2016) Dimension reduction based on the Hellinger integral. *Biometrika*, **102**, 95–106

Wedin, P.-Å. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numer. Math.*, **12**, 99–111.

Wolpert, D. (1992) Stacked generalization. *Neurl Netwrks*, **5**, 241–259.

Wright, R. M. and Switzer, P. (1971) Numerical classification applied to certain Jamaican eocene nummulitids. *Math. Geol.*, **3**, 297–311.

Yatracos, Y. G. (2002) Assessing the quality of bootstrap samples and of the bootstrap estimates obtained with finite resampling. *Statist. Probab. Lett.*, **59**, 281–292.

Yatracos, Y. G. (2013) Detecting clusters in the data from variance decompositions of its projections. *J. Classificn*, **30**, 30–55.

Yatracos, Y. G. (2017) The derivative of influence function, location breakdown point, group influence and regression residuals, plots. *Preprint*. (Available from `https://arxiv.org/pdf/1607.04384.pdf`.)

Yu, B. (2013) Stability. *Bernoulli*, **19**, 1484–1500.

Yu, Y., Li, Y.-F. and Zhou, Z.-H. (2011) Diversity regularized machine. In *Proc. Int. Jt Conf. Artificial Intelligence*, vol. 22, pp. 1603–1608. American Association for Artificial Intelligence Press.

Yu, Y., Wang, T. and Samworth, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.