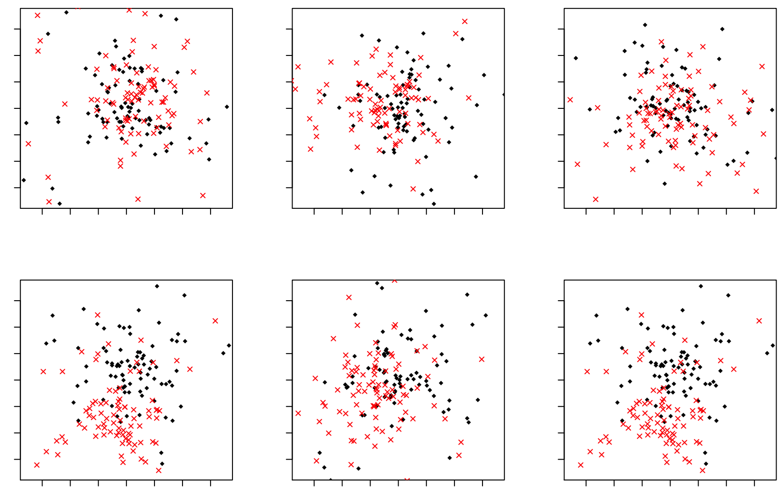
Random projection ensemble classification



Richard Samworth, University of Cambridge

Joint work with Tim Cannings

Tim Cannings

High-dimensional classification

Supervised classification problems are very frequently encountered in applications: spam filtering, fraud detection, medical diagnoses, market research,. . ..

An increasing number of modern classification problems are *high-dimensional*. Many existing techniques, e.g. LDA, may become intractable (Bickel and Levina, 2004).

Some proposals assume optimal decision boundary is linear (Friedman, 1989; Hastie et al. 1995); others assume only a few features are relevant (Fan and Fan, 2008; Tibshirani et al. 2003; Guo et al. 2007).

Random projections

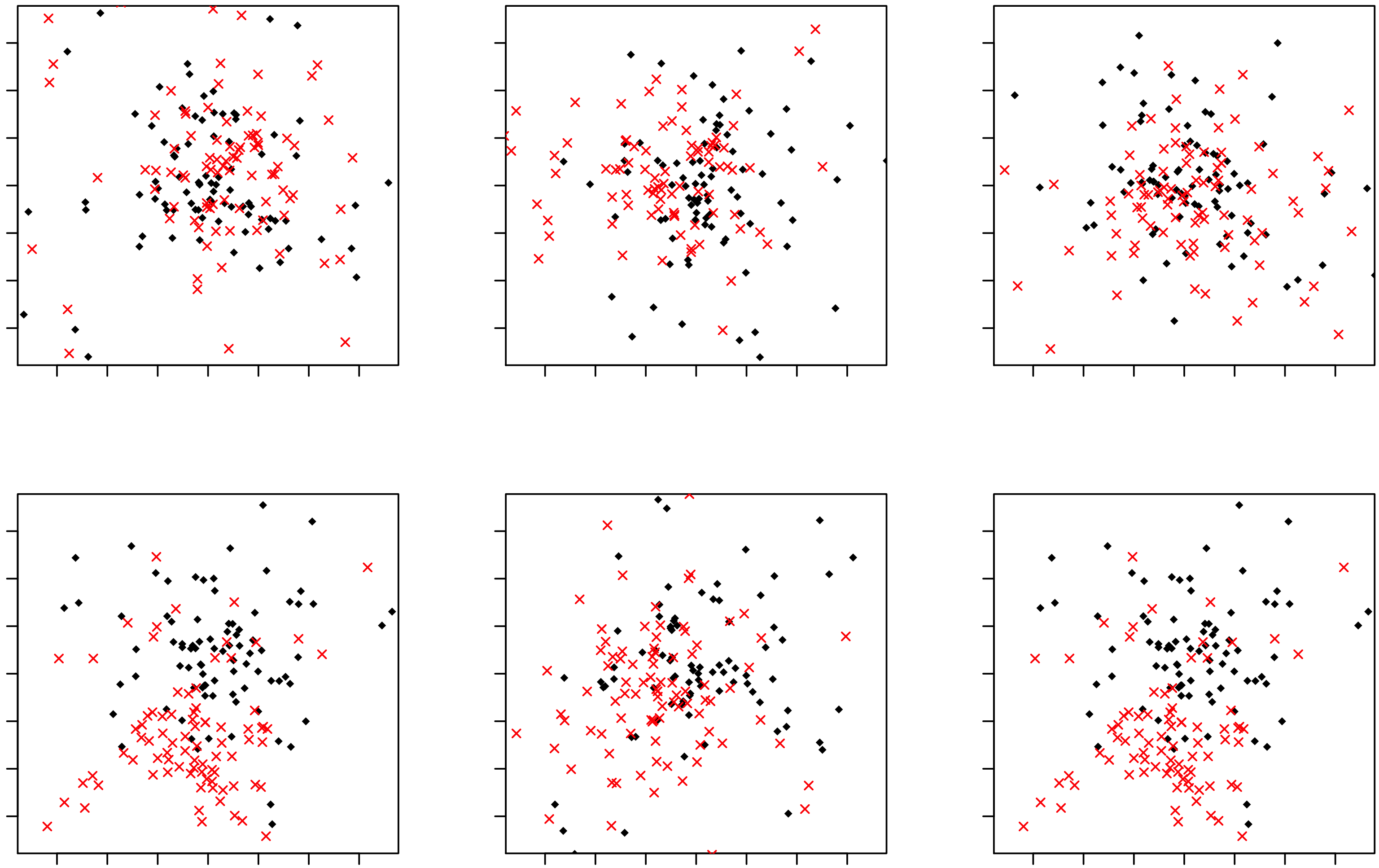
Johnson–Lindenstrauss lemma: given $x_1, \ldots, x_n \in \mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > \frac{8 \log n}{\epsilon^2}$, there exists a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ such that

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

for all $i, j = 1, \ldots, n$. Note that the lower bound on d does not depend on p.

Random projections have therefore been used successfully as a computational time saver (Durrant and Kaban, 2013; Dasgupta, 1999; McWilliams et al., 2014).

Most random projections are useless!



Setting

Suppose $(X, Y) \sim P$ **on** $\mathbb{R}^p \times \{1, 2\}$**. Let** $\pi_1 := \mathbb{P}(Y = 1)$**, and** P_r **denote the conditional distribution of** $X|Y = r$**, for** $r = 1, 2$**. Let** P_X **denote the marginal distribution of** X **and write** $\eta(x) := \mathbb{P}(Y = 1|X = x)$ **for the regression function.**

A *classifier* on $\mathbb{R}^p$ **is a measurable function** $C : \mathbb{R}^p \to \{1, 2\}$**, so we assign** $x \in \mathbb{R}^p$ **to class** $C(x)$**. The *risk*, of a classifier** C **is** $\mathcal{R}(C) := \mathbb{P}\{C(X) \neq Y\}$**, and is minimised by the *Bayes* classifier:**

$$C^{\mathrm{Bayes}}(x) = \begin{cases} 1 & \textbf{if } \eta(x) \geq 1/2; \\ 2 & \textbf{otherwise.} \end{cases}$$

Its risk is $\mathcal{R}(C^{\mathrm{Bayes}}) = \mathbb{E}[\min\{\eta(X), 1 - \eta(X)\}]$**.**

Projected data base classifier

For now, the training data $\mathcal{T}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ **are considered as fixed points in** $\mathbb{R}^p \times \{1, 2\}$**.**

Assume we have a base classifier $\hat{C}_n = \hat{C}_{n,\mathcal{T}_{n,d}}$ **that can be constructed from an arbitrary training sample** $\mathcal{T}_{n,d}$ **of size** n **in** $\mathbb{R}^d \times \{1, 2\}$**.**

Let $\mathcal{A} = \mathcal{A}_{d\times p} := \{A \in \mathbb{R}^{d\times p} : AA^T = I_{d\times d}\}$ **be the set of all** d**-dimensional projections. Given** $A \in \mathcal{A}$**, define** $z_i^A := Ax_i$ **and** $y_i^A := y_i$**, 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** $\hat{C}_n$ **is**

$$\hat{C}_n^A(x) = \hat{C}^A_{n,\mathcal{T}_n^A}(x) := \hat{C}_{n,\mathcal{T}_n^A}(Ax).$$

Random projection ensemble classifier

Let $A_1, \ldots, A_{B_1}$ **denote i.i.d. projections, independent of** (X, Y)**. Set**

$$\hat{\nu}_n^{B_1}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{\hat{C}_n^{A_{b_1}}(x)=1\}}.$$

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

$$\hat{C}_n^{\mathrm{RP}}(x) := \begin{cases} 1 & \textbf{if } \hat{\nu}_n^{B_1}(x) \geq \alpha\textbf{;} \\ 2 & \textbf{otherwise.} \end{cases}$$

Infinite-simulation version

We want to analyse $\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) := \mathbb{P}\{\hat{C}_n^{\mathrm{RP}}(X) \neq Y\}$**. Let**

$$\hat{\mu}_n(x) := \mathbb{E}\{\hat{\nu}_n^{B_1}(x)\} = \mathbb{P}\{\hat{C}_n^{A_1}(x) = 1\},$$

where the randomness comes from the random projections. Let

$$\hat{C}_n^{\mathrm{RP}^*}(x) := \begin{cases} 1 & \textbf{if } \hat{\mu}_n(x) \geq \alpha\textbf{;} \\ 2 & \textbf{otherwise.} \end{cases}$$

Asymptotic expansion

Write $G_{n,r}$ **for the distribution function of** $\hat{\mu}_n(X)|\{Y = r\}$**.**
Assume:

(A.1) $G_{n,1}$ **and** $G_{n,2}$ **are twice differentiable at** α**.**

Then

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) - \mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) = \frac{\gamma_n(\alpha)}{B_1} + o\Big(\frac{1}{B_1}\Big)$$

as $B_1 \to \infty$**, where**

$$\gamma_n(\alpha) := (1 - \alpha - [\![B_1\alpha]\!])h(\alpha) + \frac{\alpha(1-\alpha)}{2}h'(\alpha),$$

and $h(t) := \pi_1 g_{n,1}(t) - \pi_2 g_{n,2}(t)$**.**

Infinite-simulation classifier test error

Define the test error of $\hat{C}_n^A$ by

$$\mathcal{L}_n^A := \int_{\mathbb{R}^p \times \{1,2\}} \mathbb{1}_{\{\hat{C}_n^A(x) \neq y\}} \, dP(x,y).$$

Then, with no assumptions $\mathcal{T}_n$, the distribution P or on the distribution of the individual projections,

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}^*}) - \mathcal{R}(C^{\mathrm{Bayes}}) \leq \frac{1}{\min(\alpha, 1-\alpha)} \{\mathbb{E}(\mathcal{L}_n^{A_1}) - \mathcal{R}(C^{\mathrm{Bayes}})\}.$$

Choosing good random projections

Let $\hat{L}_n^A = \hat{L}_n^A(z_1^A, y_1^A, \ldots, z_n^A, y_n^A)$ **be an estimator of** $\mathcal{L}_n^A$ **taking values in** $\{0, 1/n, \ldots, 1\}$**. For** $B_1, B_2 \in \mathbb{N}$**, let** $\{A_{b_1,b_2} : b_1 = 1, \ldots, B_1, b_2 = 1, \ldots, B_2\}$ **denote independent projections, independent of** (X, Y)**, from Haar measure on** $\mathcal{A}$**. For** $b_1 = 1, \ldots, B_1$**, let**

$$b_2^*(b_1) := \operatorname*{sargmin}_{b_2 \in \{1,\ldots,B_2\}} \hat{L}_n^{A_{b_1,b_2}}. \tag{1}$$

We now set $A_{b_1} := A_{b_1, b_2^*(b_1)}$**, and consider** $\hat{C}_n^{\mathrm{RP}}$ **using the independent projections** $A_1, \ldots, A_{B_1}$**.**

The induced Bayes classifier

For $z \in \mathbb{R}^d$ and $A \in \mathcal{A}$ define $\eta^A(z) := \mathbb{P}(Y = 1|AX = z)$. The induced Bayes classifier, which is the optimal classifier knowing only the distribution of (AX, Y), is

$$C^{A-\mathrm{Bayes}}(z) := \begin{cases} 1 & \textbf{if } \eta^A(z) \geq 1/2\textbf{;} \\ 2 & \textbf{otherwise.} \end{cases}$$

Its risk is

$$\mathcal{R}^{A-\mathrm{Bayes}} := \int_{\mathbb{R}^p \times \{1,2\}} \mathbb{1}_{\{C^{A-\mathrm{Bayes}}(Ax) \neq y\}} \, dP(x, y).$$

Two further conditions

Let

$$\hat{L}_n^* := \min_{A \in \mathcal{A}} \hat{L}_n^A$$

denote the optimal test error estimate over all projections. For $j = 0, 1, \ldots, \lfloor n(1 - \hat{L}_n^*) \rfloor$**, let**

$$\beta_n(j) := \mathbb{P}\big(\hat{L}_n^A \leq \hat{L}_n^* + j/n\big),$$

where $A \sim \mathrm{Haar}(\mathcal{A})$**. We will assume:**

(A.2) There exist $\beta_0 \in (0,1)$ **and** $\beta, \rho > 0$ **such that**

$$\beta_n(j) \geq \beta_0 + \frac{\beta j^\rho}{n^\rho}$$

for $j \in \Big\{0, 1, \ldots, \Big\lfloor n\Big(\frac{\log^2 B_2}{\beta B_2}\Big)^{1/\rho}\Big\rfloor + 1\Big\}$**.**

Bayes classifier condition

(A.3) There exists a projection $A^* \in \mathcal{A}$ such that

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

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

If the Bayes decision boundary is a hyperplane, then (A.3) holds with $d = 1$. Moreover, if Y is independent of X given A^*X, then (A.3) holds.

Final bound

Assume (A.1), (A.2) and (A.3). Then

$$\mathcal{L}(\hat{C}_n^{\mathrm{RP}}) - \mathcal{R}(C^{\mathrm{Bayes}}) \leq \frac{\mathcal{L}_n^{A^*} - \mathcal{R}^{A^*-\mathrm{Bayes}}}{\min(\alpha, 1-\alpha)} + \frac{\epsilon_n - \epsilon_n^{A^*}}{\min(\alpha, 1-\alpha)} + \frac{\gamma_n(\alpha)}{B_1}\{1 + o(1)\} + g_n(B_2; \beta_0, \beta, \rho)$$

as $B_1 \to \infty$**, where** $\epsilon_n = \epsilon_n^{(B_2)} := \mathbb{E}(\mathcal{L}_n^{A_1} - \hat{L}_n^{A_1})$**,**
$\epsilon_n^{A^*} := \mathcal{L}_n^{A^*} - \hat{L}_n^{A^*}$ **and**

$$g_n(B_2; \beta_0, \beta, \rho) := \frac{(1-\beta_0)^{B_2}}{\min(\alpha, 1-\alpha)} \left\{ \frac{1}{n} + \frac{(1-\beta_0)^{1/\rho}\Gamma(\frac{1+\rho}{\rho})}{B_2^{1/\rho}\beta^{1/\rho}} + e^{-\frac{\log^2 B_2}{1-\beta_0}} \right\}.$$

Choice of base classifier: LDA

We now regard $\mathcal{T}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ **as independent random pairs with distribution** P**. If** $X|Y = r \sim N_p(\mu_r, \Sigma)$**, then**

$$\mathrm{sgn}\{\eta(x) - 1/2\} = \mathrm{sgn}\biggl\{\log\frac{\pi_1}{\pi_2} + \Bigl(x - \frac{\mu_1 + \mu_2}{2}\Bigr)^T \Sigma^{-1}(\mu_1 - \mu_2)\biggr\},$$

so (A.3) holds with $d = 1$ **and** $A^* = \frac{(\mu_1 - \mu_2)^T \Sigma^{-1}}{\|\Sigma^{-1}(\mu_1 - \mu_2)\|}$**. If** $Y_1 = \ldots = Y_{n_1} = 1$ **and** $Y_{n_1+1} = \ldots = Y_n = 2$**, then**

$$\mathbb{E}(\mathcal{L}_n^{A^*}) - \mathcal{R}^{A^*-\mathrm{Bayes}} = \frac{d}{n}\phi\Bigl(-\frac{\Delta}{2}\Bigr)\biggl\{\frac{\Delta}{4} + \frac{d-1}{d\Delta}\biggr\}\{1 + O(n^{-1})\}$$

where $\Delta := \|\Sigma^{-1/2}(\mu_1 - \mu_2)\| = \|(\Sigma^{A^*})^{-1/2}(\mu_1^{A^*} - \mu_2^{A^*})\|$ **when** $n_1 = n_2 = n/2$ **(Okamoto, 1963).**

Controlling $\epsilon_n^{A^*}$ and ϵ_n

Consider the resubstitution estimate

$$\hat{L}_n^A := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat{C}_n^{A-\mathrm{LDA}}(X_i)\neq Y_i\}}.$$

We have the Vapnik–Chervonenkis bound:

$$\sup_{A\in\mathcal{A}} \mathbb{P}(|\mathcal{L}_n^A - \hat{L}_n^A| > \epsilon) \leq 8n^d e^{-n\epsilon^2/32}$$

(Devroye and Wagner, 1976). **We deduce that**

$$\mathbb{E}(|\epsilon_n^{A^*}|) \leq 8\sqrt{\frac{d\log n + 3\log 2 + 1}{2n}}$$

$$\mathbb{E}(|\epsilon_n|) \leq 8\sqrt{\frac{d\log n + 3\log 2 + \log B_2 + 1}{2n}}.$$

k-nearest neighbour classifier

Under regularity conditions,

$$\mathbb{E}(\mathcal{L}_n^{A^*}) - \mathcal{R}(C^{A^*-\mathrm{Bayes}}) = O(n^{-4/(d+4)})$$

(Hall, Park and S., 2008, S., 2012). Consider $\hat{L}_n^A := n^{-1}\sum_{i=1}^n \mathbb{1}_{\{\hat{C}_{n,i}^A(X_i)\neq Y_i\}}$**, where** $\hat{C}_{n,i}^A$ **is trained on** $\mathcal{T}_n^A \setminus \{X_i^A, Y_i^A\}$**. Then**

$$\mathbb{E}(|\epsilon_n^{A^*}|) \leq \Big(\frac{1}{n} + \frac{24k^{1/2}}{n\sqrt{2\pi}}\Big)^{1/2} \leq \frac{1}{n^{1/2}} + \frac{2\sqrt{3}k^{1/4}}{n^{1/2}\sqrt{\pi}}$$

$$\mathbb{E}(|\epsilon_n|) \leq 3\{4(3^d+1)\}^{1/3}\bigg\{\frac{k(1+\log B_2 + 3\log 2)}{n}\bigg\}^{1/3}.$$

Practical considerations: choice of α

Since $\mathcal{L}(\hat{C}_n^{\mathrm{RP}*}) = \pi_1 G_{n,1}(\alpha) + \pi_2\{1 - G_{n,2}(\alpha)\}$**, we have the 'oracle' choice**

$$\alpha^* \in \operatorname*{argmin}_{\alpha' \in [0,1]} \bigl[\pi_1 G_{n,1}(\alpha') + \pi_2\{1 - G_{n,2}(\alpha')\}\bigr].$$

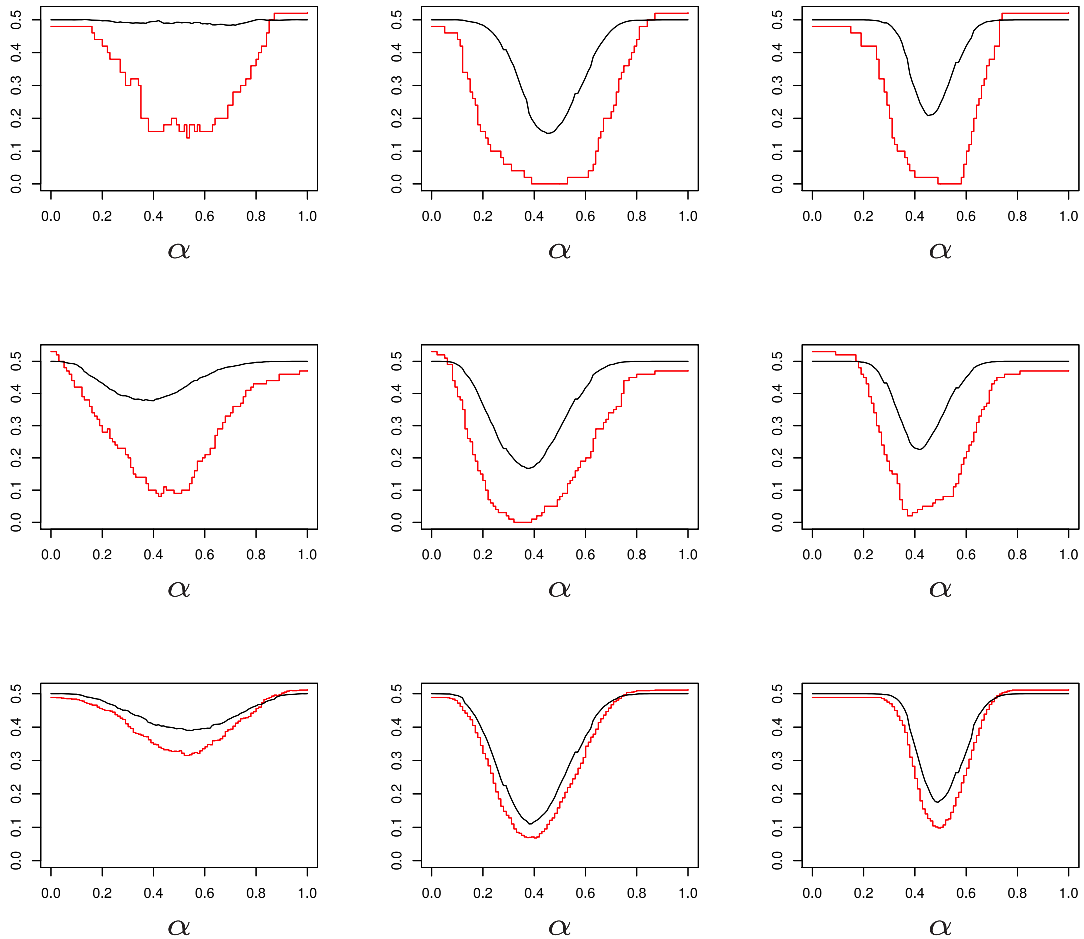
We can estimate $G_{n,r}$ **using**

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

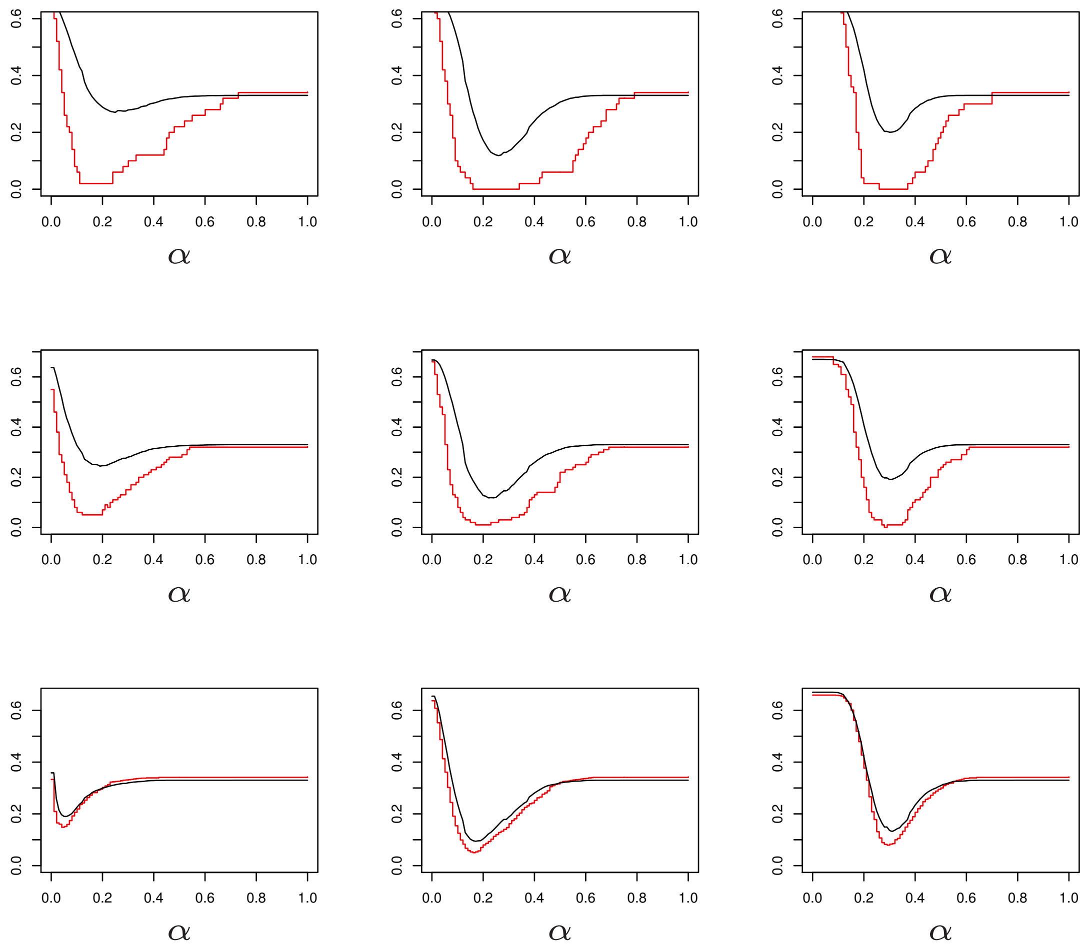
for $r = 1, 2$**, and set**

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha' \in [0,1]} \bigl[\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_2\{1 - \hat{G}_{n,2}(\alpha')\}\bigr].$$

Choice of α



Choice of α



Simulation results: multi-modal features

P_1: $\frac{1}{2}N_p(\mu, I_p) + \frac{1}{2}N_p(-\mu, I_p)$; P_2: **Indep. comp., 5 Cauchy, $p - 5$ $N(0, 1)$, $\mu = (1, \ldots, 1, 0, \ldots, 0)$ with 5 non-zeros.**

	$\pi_1 = 0.5$, **BR = 11.58**		
n	50	100	200
RP-QDA$_5$	29.93	**24.83**	**22.20**
RP-$k$nn$_5$	**29.36**	26.29	23.38
QDA	N/A	N/A	27.58
Random Forest	40.49	33.51	25.02
Linear SVM	48.66	49.31	48.87
Radial SVM	48.87	49.66	48.18
PenLDA	48.34	48.97	49.21
NSC	47.67	47.69	47.83
SCRDA	45.44	44.86	43.27

Simulation results: No (A.3)

P_1: **independent Laplace;** $P_2 : N_p(\mu, I_{p \times p})$, $\mu = \frac{1}{8}(1, \ldots, 1)$.

	$\pi_1 = 0.33$, **BR = 4.09**		
n	50	100	200
RP-QDA$_2$	**17.64**	13.37	11.88
RP-QDA$_5$	18.06	**12.86**	**10.64**
QDA	N/A	N/A	33.05
Random Forest	31.65	28.21	22.92
Linear SVM	36.50	35.84	31.82
Radial SVM	32.03	30.48	22.27
PenLDA	33.19	32.61	31.31
NSC	31.76	31.13	31.65
SCRDA	33.56	32.52	31.94
IR	35.04	36.26	36.48

Musk molecule dataset

1016 musk, 5581 non-musk molecules, $p = 166$ features

n	50	100	200
RP-QDA$_5$	14.70	12.72	9.93
RP-$k$nn$_5$	**13.88**	**10.96**	**8.67**
QDA	N/A	N/A	N/A
$k$nn	16.22	14.41	11.14
Random Forest	14.40	13.18	10.67
Linear SVM	16.49	13.91	10.39
Radial SVM	15.27	15.25	15.21
PenLDA	29.57	27.76	27.15
NSC	16.41	15.45	15.19
SCRDA	15.69	16.40	15.14
IR	32.22	30.83	30.58

Extensions: sample splitting

Split the sample $\mathcal{T}_n$ into $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$, and use

$$\hat{L}^A_{n^{(1)},n^{(2)}} := \frac{1}{n^{(2)}} \sum_{(X_i,Y_i)\in\mathcal{T}_{n,2}} \mathbb{1}_{\{\hat{C}^A_{n^{(1)},\mathcal{T}^A_{n,1}}(X_i)\neq Y_i\}}$$

to estimate the test error $\mathcal{L}^A_{n^{(1)},1}$ based on the training data $\mathcal{T}_{n,1}$. By Hoeffding's inequality,

$$\sup_{A\in\mathcal{A}} \mathbb{P}\bigl\{|\mathcal{L}^A_{n^{(1)},1} - \hat{L}^A_{n^{(1)},n^{(2)}}| \geq \epsilon \bigm| \mathcal{T}_{n,1}\bigr\} \leq 2e^{-2n^{(2)}\epsilon^2}.$$

It then follows that

$$\mathbb{E}\bigl(|\mathcal{L}^{A_1}_{n^{(1)}} - \hat{L}^{A_1}_{n^{(1)},n^{(2)}}| \bigm| \mathcal{T}_{n,1}\bigr) \leq \Bigl(\frac{1+\log 2+\log B_2}{2n^{(2)}}\Bigr)^{1/2}.$$

Extensions: Multiclass problems

For $K > 2$ classes, we can let

$$\hat{\nu}_{n,r}^{B_1}(x) := \frac{1}{B_1}\sum_{b_1=1}^{B_1} \mathbb{1}_{\{\hat{C}_n^{A_{b_1}}(x)=r\}}$$

for $r = 1, \ldots, K$. Given $\alpha_1, \ldots, \alpha_K > 0$ with $\sum_{r=1}^{K} \alpha_r = 1$, we can then define

$$\hat{C}_n^{\mathrm{RP}}(x) := \operatorname*{sargmax}_{r=1,\ldots,K}\{\alpha_r \hat{\nu}_{n,r}^{B_1}(x)\}.$$

The choice of $\alpha_1, \ldots, \alpha_K$ is analogous to the choice of α in the case $K = 2$.

Extensions: Ultrahigh dimensions

When p is huge, it may be too time-consuming to generate enough random projections.

We can instead restrict A to be axis-aligned, so that each row of A consists of a single non-zero component, equal to 1, and $p - 1$ zero components. Here, there are only $\binom{p}{d} \leq p^d/d!$ choices.

Corresponding theory can be obtained provided that the projection A^* in (A.3) is axis-aligned.

References

- 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.

- Cannings, T. I. and Samworth, R. J. (2015). Random projection ensemble classification. `arxiv 1504.04595`.

- Dasgupta, S. (1999). Learning mixtures of Gaussians. *Proc. 40th Annual Symposium on Foundations of Computer Science*, 634–644.

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

- Durrant, R. J. and Kaban, A. (2013). Sharp generalization error bounds for randomly-projected classifiers. *Journal of Machine Learning Research-Proceedings*, 28, 693–701.

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

- **Friedman, J. (1989). Regularised discriminant analysis. *J. Amer. Statist. Assoc.*, 84, 165–175.**

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

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

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

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

- **Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discrminant function.**

***Ann. Math. Statist.*, 34, 1286–1301.**

- **Tibshirani, R., Hastie, T., Narisimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centriods, with applications to DNA microarrays. *Statist. Science*, 18, 104–117.**
- **Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discrminant function. *Ann. Math. Statist.*, 34, 1286–1301.**
- **Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, 40, 2733–2763.**