Biometrika (2012), **xx**, x, *pp*. 1–11

Printed in Great Britain

Supplementary Material for Componentwise classification and clustering of functional data

BY A. DELAIGLE, P. HALL

Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

A.Delaigle@ms.unimelb.edu.au halpstat@ms.unimelb.edu.au

AND N. BATHIA

Jump Trading, 600 West Chicago Avenue, Chicago, Illinois 60654, United States of America

nbathia@jumptrading.com

1. PROCEDURES FOR BREAKING TIES

Since $\hat{\text{err}}_r$ can take at most $n+1$ different values, its minimum is not always unique. For the linear and quadratic discriminant methods, as well as for the nonparametric Bayes procedure, we suggest breaking ties as follows. First, note that these three methods are based on the Bayes rule. They assign x to population 0, i.e. in the notation above, $J(x, \mathcal{D} \,|\, t_{(r)}) = 0$, if

$$\pi_0 \, \tilde{f}_0(x \,|\, t_{(r)}) > \pi_1 \, \tilde{f}_1(x \,|\, t_{(r)}), \tag{1}$$

and to population 1, i.e. $J(x, \mathcal{D} \,|\, t_{(r)}) = 1$, otherwise. For the nonparametric Bayes rule, $\tilde{f}_k(x \,|\, t_{(r)}) = \hat{f}_k(x \,|\, t_{(r)})$; for Fisher's linear and quadratic discriminant methods, the $\tilde{f}_k(x \,|\, t_{(r)})$s are r-variate normal densities with means $\bar{\mathrm{X}}_k(t_{(r)})$ and covariance matrix $\hat{\Sigma}(t_{(r)})$ for linear discriminant, or covariance matrices $\hat{\Sigma}_k(t_{(r)})$ for quadratic discriminant. In the case of ties, we choose among them the vector $t_{(r)}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}|\breve{f}_0(X_i\,|\,t_{(r)})-\breve{f}_1(X_i\,|\,t_{(r)})|/\max\{\breve{f}_0(X_i\,|\,t_{(r)}),\breve{f}_1(X_i\,|\,t_{(r)})\}, \tag{2}$$

where, for $k=0,1$, $\breve{f}_k$ denotes an estimator of f_k. In the linear and quadratic discriminant cases, we took $\breve{f}_k=\tilde{f}_k$ defined above; in the nonparametric case we took $\breve{f}_k=\tilde{f}_{k,i}$, where $\tilde{f}_{k,i}$ denotes the estimator of f_k constructed without using X_i. The criterion at (2) is an empirical mean distance between $\tilde{f}_0$ and $\tilde{f}_1$, relative to the magnitude of $\tilde{f}_0$ and $\tilde{f}_1$.

For the classifier based on nonparametric regression, we break ties by choosing among them the one that minimizes the leave-one-out absolute error of the regression fit, $\sum_{i=1}^{n}|I_i-\hat{g}_i\big(X_i\mid t_{(r)}\big)|$, where $\hat{g}_i$ denotes the estimator of g constructed without using X_i. For the classifier based on logistic regression, we choose among ties the one that minimizes the Akaike information criterion; if there are still ties with this criterion, we create a noisy version of the training data X_i by adding to each component a normal random variable with mean zero and variance $0{\cdot}1$ times the empirical variance of the component, and then break the ties by calculating the estimator of error rate from these perturbed data, followed, if necessary, by the Akaike information criterion.

2. ADDITIONAL SIMULATION RESULTS

2·1. *Comparison of nonparametric Bayes and regression-based classifiers*

As indicated in §4.2 of the paper, the nonparametric Bayes classifier gave results similar to the nonparametric regression-based one. This is illustrated in Fig. 1 below, which compares the results of the nonparametric Bayes and regression classifiers for the three datasets considered in the paper, and for training samples of sizes $n=30$, 50 and 100. The boxplots were constructed from 200 Monte Carlo replications, as in the paper.

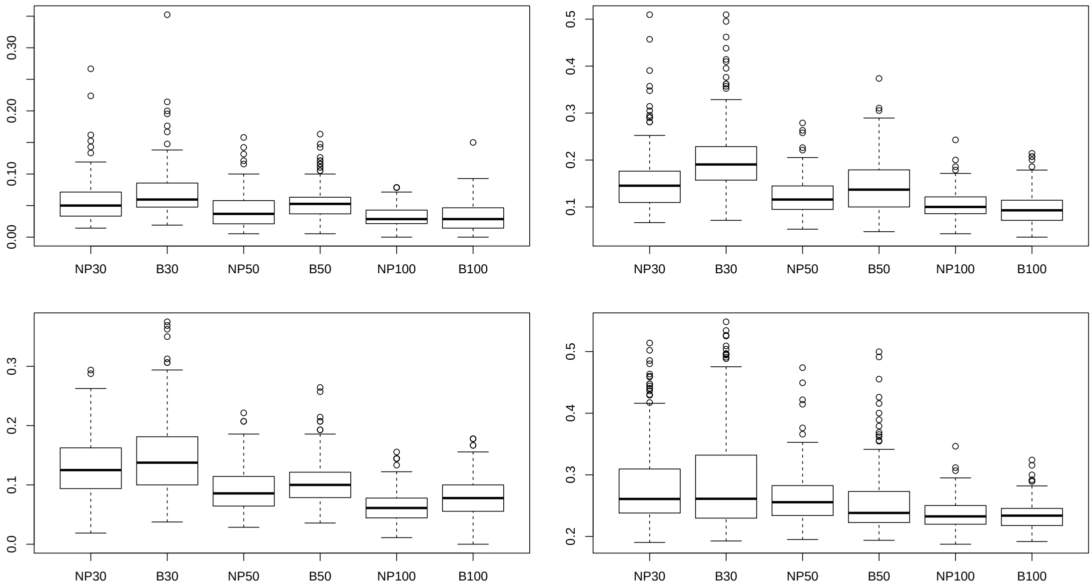

Fig. 1. Comparison of results for the Bayes (B) method and the nonparametric regression (NP) procedure, for training samples of sizes $n = 30$, 50 and 100. Top: Tecator data, where the left panel shows case I and the right panel shows case II. Bottom left: rainfall data, bottom right: phoneme data.

We can see that overall the regression-based classifier outperformed its Bayes counterpart for $n \leq 50$, but the Bayes classifier improves when $n = 100$. In part, this can be explained by the fact that the regression-based classifier requires only one bandwidth, constructed from the entire sample, whereas the Bayes classifier requires two bandwidths (one for each group), each constructed from observations in one group only. For n small, the groups can be of rather low size, which makes the bandwidth choice too variable, but when n is larger, the group sizes are adequate for these bandwidths to be reliable, and hence for the Bayes classifier to work well.

2·2. *Number of selected points*

Fig. 2 shows the frequency at which $k = 1, \ldots, 5$ points were selected over 200 Monte Carlo simulations, for training samples of sizes $n = 30$, 50 and 100 and for each of the four examples considered in our numerical work.

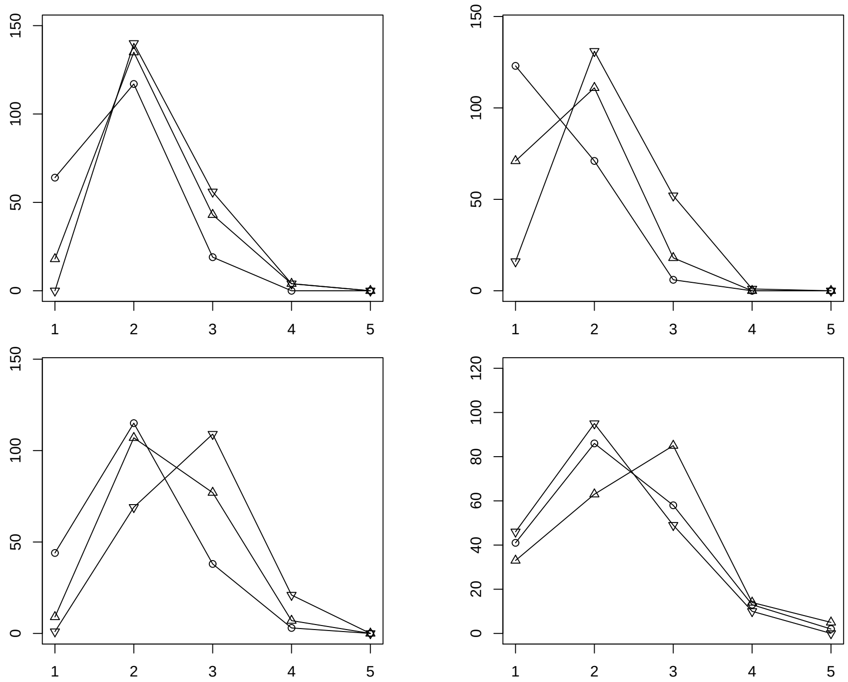

Fig. 2. Number of selected points. The graphs show the frequency at which $k = 1, \ldots, 5$ points were selected over 200 Monte Carlo simulations, for training samples of size $n = 30$ ($\circ$), 50 ($\triangle$) and 100 ($\bigtriangledown$). The graphs are for the rainfall data (top left), the Tecator data, case I (top right), the Tecator data, case II (bottom left) and the phoneme data (bottom right).

2·3. *Effect of* ρ

As mentioned below equation (3) in §2.2 of the paper, our method is not very sensitive to choice of the threshold. To illustrate this point, in Figure 3 below we show, for training samples of sizes $n = 30$, 50 and 100, boxplots of the classification error rates, calculated from 200 Monte Carlo replications, and obtained when applying our method with nonparametric regression-based, logistic, linear discriminant and quadratic discriminant classifiers when $\rho = 0$, $0{\cdot}1$ and $0{\cdot}2$. We can see that the results for $\rho = 0$ and $\rho = 0{\cdot}1$ are almost identical, and the results for $\rho = 0{\cdot}2$ do not differ much.

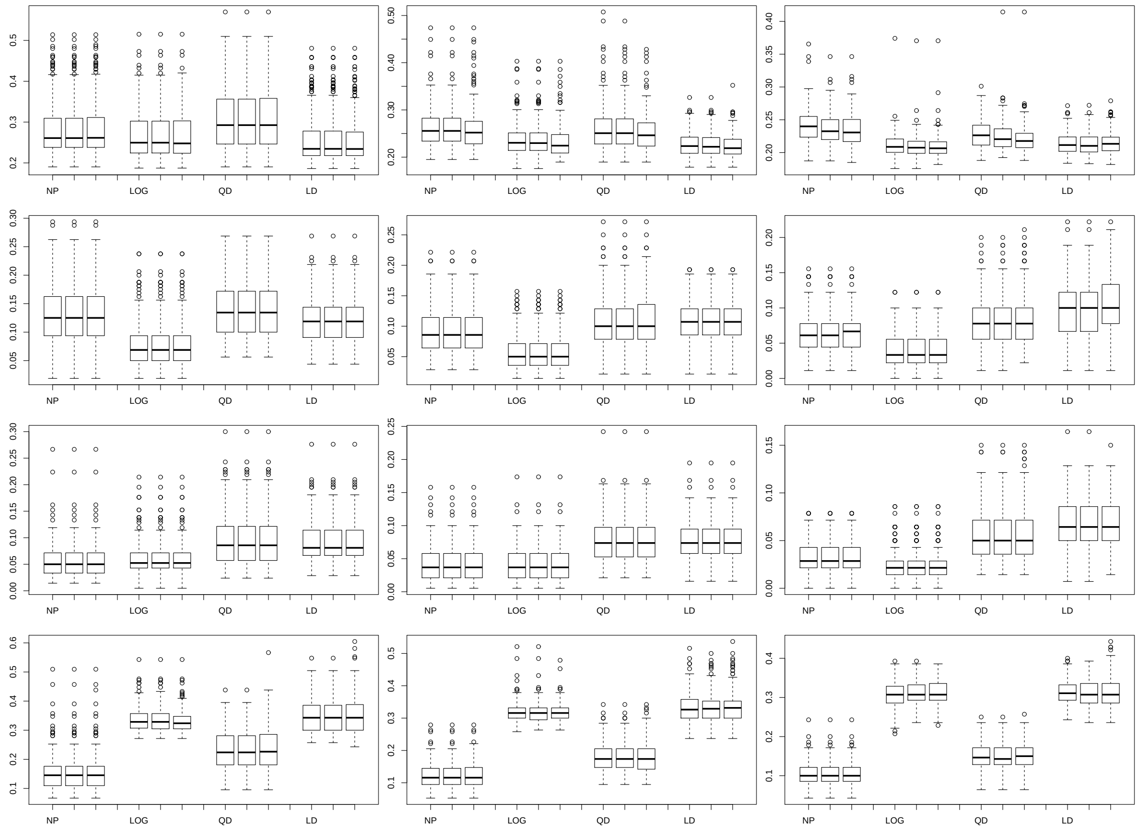

Fig. 3. Effect of ρ on the nonparametric regression-based method combined with our approach (NP), the logistic regression methods combined with our approach (LOG), the linear discriminant method combined with our approach (LD) and the quadratic discriminant method (QD). In each group of three boxplots, the first is for $\rho = 0$, the second for $\rho = 0{\cdot}1$ and the third for $\rho = 0{\cdot}2$. The first column is for training samples of size $n = 30$, the second column is for $n = 50$, and the third column is for $n = 100$. Rows 1 to 4 show, respectively the Phoneme, Rain, Tecator, case I and Tecator, case II data.

2·4. *Tables*

Table 1 shows means and standard deviations of the percentage of misclassified observations calculated from $M = 200$ Monte Carlo replications for each of the data sets considered in our numerical work, and for training samples of sizes $n = 30$, 50 and 100.

Table 1. *Mean (standard deviation) of the percentage of misclassified observations calculated from $M = 200$ Monte Carlo replications from the rainfall data (Rain), the Tecator data (Tec), cases I and II, and the phoneme data (Phon). The results are shown for the nonparametric regression-based methods combined with our approach (NP), with principal components (NPC) or with partial least-squares (NPLS), the boosting version of NP (NPb), the logistic regression methods combined with our approach (LOG), with partial least-squares (LOGPLS) and with boosting (LOGb), the linear discriminant method combined with our approach (LD) and with partial least-squares (LDPLS), and the quadratic discriminant method (QD).*

Data	n	NPC	NP	NPb	NPPLS	LOG	LOGb	LOGPLS	QD	LD	LDPLS
Rain	30	13 (5.0)	13 (5.5)	11 (4.8)	10 (4.6)	8.0 (4.2)	8.0 (4.2)	8.8 (4.0)	14 (5.1)	12 (3.7)	11 (3.8)
	50	9.2 (4.0)	9.2 (3.8)	8.3 (3.4)	8.6 (3.7)	5.5 (3.0)	5.5 (3.0)	8.1 (3.4)	11 (4.4)	11 (3.1)	9.7 (3.5)
	100	5.9 (2.8)	6.4 (2.7)	5.2 (2.3)	8.0 (3.3)	4.3 (2.5)	4.2 (2.5)	7.4 (2.9)	8.2 (3.5)	9.9 (3.7)	9.1 (3.4)
Tec I	30	7.0 (3.3)	5.6 (3.3)	5.3 (3.1)	5.5 (2.4)	6.0 (3.3)	6.0 (3.3)	5.9 (3.1)	9.6 (5.1)	9.1 (3.9)	7.1 (3.4)
	50	5.8 (1.9)	4.2 (2.5)	4.1 (2.4)	4.5 (1.5)	4.2 (2.6)	4.2 (2.6)	4.7 (1.9)	7.7 (3.4)	7.7 (3.1)	6.5 (2.3)
	100	5.4 (1.6)	3.3 (1.7)	3.2 (1.7)	4.0 (1.5)	2.2 (1.6)	2.2 (1.6)	4.0 (1.5)	5.5 (2.6)	6.6 (2.7)	6.5 (2.3)
Tec II	30	22 (6.5)	15 (6.5)	15 (6.2)	33 (4.4)	34 (4.4)	26 (6.8)	33 (4.4)	23 (6.5)	35 (5.8)	35 (5.6)
	50	19 (4.2)	12 (3.9)	12 (3.7)	33 (3.7)	32 (3.3)	20 (5.2)	32 (3.4)	18 (4.6)	33 (4.7)	34 (5.0)
	100	16 (3.5)	10 (2.9)	10 (3.0)	33 (3.7)	31 (3.0)	14 (2.8)	30 (3.7)	15 (3.6)	31 (3.4)	32 (3.7)
Phon	30	33 (4.9)	28 (6.9)	28 (6.4)	27 (4.6)	27 (6.2)	27 (5.9)	27 (3.7)	31 (7.8)	26 (6.1)	27 (3.4)
	50	31 (4.6)	26 (4.2)	26 (3.9)	25 (3.8)	24 (3.6)	24 (3.3)	24 (2.8)	26 (4.8)	23 (2.6)	25 (3.1)
	100	30 (4.3)	24 (2.4)	24 (2.3)	23 (2.5)	21 (1.9)	21 (1.6)	22 (2.0)	23 (2.4)	21 (1.7)	23 (2.7)

3. TECHNICAL ARGUMENTS

Proof of Theorem 1. Let $\mathcal{A}$, depending on n, represent a lattice in $\mathcal{I}_r$ of edge width n^{-B} in each of the r components, for some $B > 0$. For any $t_{(r)} \in \mathcal{I}_r$, let $t^*_{(r)}$ be the element of $\mathcal{A}$ that is nearest to $t_{(r)} \in \mathcal{I}_r$. Then $\sup_{t_{(r)} \in \mathcal{I}_r} \|t_{(r)} - t^*_{(r)}\| = O(n^{-B})$. In the arguments below, B can be chosen arbitrarily large.

Step 1: Part (i) of the theorem

Step 1.1. Here we prove that

$$\sup_{t_{(r)} \in \mathcal{J}_r(c)} |\hat{\text{err}}_r(t_{(r)}) - \text{err}_r(t^*_{(r)})| = o_P(1). \tag{3}$$

For random variables $R_{i,1}(t_{(r)})$, $R_{i,2}(t_{(r)})$ and $R_3(t_{(r)})$ we can write:

$$\begin{aligned}
\hat{\text{err}}_r(t_{(r)}) &= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\pi_k \hat{f}_k(X_i \,|\, t_{(r)}) < \pi_{1-k} \hat{f}_{1-k,n_{1-k}}(X_i \,|\, t_{(r)})\Big\} \\
&= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\pi_k f_k(X_i \,|\, t_{(r)}) < \pi_{1-k} f_{1-k}(X_i \,|\, t_{(r)}) + R_{i,1}(t_{(r)})\Big\} \\
&= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\pi_k f_k(X_i \,|\, t^*_{(r)}) < \pi_{1-k} f_{1-k}(X_i \,|\, t^*_{(r)}) + R_{i,2}(t_{(r)})\Big\} \\
&= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\pi_k f_k(X_i \,|\, t^*_{(r)}) < \pi_{1-k} f_{1-k}(X_i \,|\, t^*_{(r)})\Big\} + R_3(t_{(r)}) \qquad (4) \\
&\equiv \tilde{\text{err}}(t^*_{(r)}) + R_3(t_{(r)}) = \text{err}_r(t^*_{(r)}) + R_3(t_{(r)}) + R_4(t^*_{(r)}),
\end{aligned}$$

where

$$\tilde{\text{err}}_r(t^*_{(r)}) = \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\pi_k f_k(X_i \,|\, t^*_{(r)}) < \pi_{1-k} f_{1-k}(X_i \,|\, t^*_{(r)})\Big\}.$$

For all $\epsilon > 0$ and all $k \geq 1$,

$$\begin{aligned}
\text{pr}\Big\{\sup_{t^*_{(r)} \in \mathcal{A}} |R_4(t^*_{(r)})| > \epsilon\Big\} &= \text{pr}\Big\{\sup_{t^*_{(r)} \in \mathcal{A}} |\tilde{\text{err}}(t^*_{(r)}) - \text{err}(t^*_{(r)})| > \epsilon\Big\} \\
&= \text{pr}\Big[\sup_{t^*_{(r)} \in \mathcal{A}} |\tilde{\text{err}}(t^*_{(r)}) - E\{\tilde{\text{err}}(t^*_{(r)})\}| > \epsilon\Big] \\
&\leq c_1 n^B \sup_{t^*_{(r)} \in \mathcal{A}} \text{pr}\Big[|\tilde{\text{err}}(t^*_{(r)}) - E\{\tilde{\text{err}}(t^*_{(r)})\}| > \epsilon\Big] = O\big(n^{-C_2}\big),
\end{aligned}$$

where $c_1 > 0$ is a finite constant, and for all $C_2 > 0$, where the $O(n^{-C_2})$ bound follows using Bernstein's inequality. Therefore $\sup_{t^*_{(r)} \in \mathcal{A}} |R_4(t^*_{(r)})| \to 0$ in probability. Result (3) is a consequence of this property and the next two results, which we derive next: for $\ell = 1, 2$,

$$\sup_{i, t_{(r)} \in \mathcal{J}_r(c)} |R_{i,\ell}(t_{(r)})| = o_P(1), \tag{5}$$

$$\sup_{t_{(r)} \in \mathcal{J}_r(c)} |R_3(t_{(r)})| = o_P(1). \tag{6}$$

Under the conditions of the theorem, standard arguments based on approximating $\hat{f}_k(x \,|\, t_{(r)}) - E\{\hat{f}_k(x \,|\, t_{(r)})\}$ and $E\{\hat{f}_k(x \,|\, t_{(r)})\} - f_k(x \,|\, t_{(r)})$, for values x and $t_{(r)}$ on lattices of polynomial denseness, can be used to prove that, for $k = 0, 1$,

$$\sup_{x \in \mathbb{R}^r} \sup_{t_{(r)} \in \mathcal{J}_r(c)} |\hat{f}_k(x \,|\, t_{(r)}) - f_k(x \,|\, t_{(r)})| = o_P(1). \tag{7}$$

Therefore,

$$|R_{i,1}(t_{(r)})| \le \sum_{k=0}^{1} \pi_k |f_k(X_i \,|\, t_{(r)}) - \hat{f}_k(X_i \,|\, t_{(r)})| \to 0$$

in probability, uniformly in $t_{(r)} \in \mathcal{J}_r(c)$. This proves that (5) holds for $\ell = 1$.

To show that (5) holds for $\ell = 2$, note that

$$\begin{aligned} |R_{i,2}(t_{(r)})| &\le |R_{i,1}(t_{(r)})| + \sum_{k=0}^{1} \pi_k |f_k(X_i \,|\, t_{(r)}) - f_k(X_i \,|\, t^*_{(r)})| \\ &\le |R_{i,1}(t_{(r)})| + \sup_{x \in \mathbb{R}^r} \max_{k=0,1} \sup_{t_{(r)} \in \mathcal{J}_r(c)} |f_k(x \,|\, t_{(r)}) - f_k(x \,|\, t^*_{(r)})|. \end{aligned}$$

These bounds, (5) for $\ell = 1$, and Condition A(e) imply that (5) holds for $\ell = 2$. To prove (6), recall from the definition of $R_3(t_{(r)})$ at (4), and (5) for $\ell = 2$, that, for all $\eta > 0$, the following result holds uniformly in $t_{(r)} \in \mathcal{J}(c)$:

$$R_3(t_{(r)}) \le R_5(t_{(r)}, \eta) + \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\{|R_{i,2}(t_{(r)})| > \eta\} = R_5(t_{(r)}, \eta) + o_p(1), \tag{8}$$

where

$$\begin{aligned} R_5(t_{(r)}, \eta) &= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} I\Big\{\big|\pi_k f_k(X_i \,|\, t^*_{(r)}) - \pi_{1-k} f_{1-k}(X_i \,|\, t^*_{(r)})\big| \le \eta\Big\} \\ &= \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} (1 - E) I\Big\{\big|\pi_k f_k(X_i \,|\, t^*_{(r)}) - \pi_{1-k} f_{1-k}(X_i \,|\, t^*_{(r)})\big| \le \eta\Big\} \\ &\quad + \frac{1}{n} \sum_{k=0}^{1} n_k \mathrm{pr}_k \Big\{\big|\pi_k f_k(X \,|\, t^*_{(r)}) - \pi_{1-k} f_{1-k}(X \,|\, t^*_{(r)})\big| \le \eta\Big\}. \end{aligned}$$

Bernstein's inequality can be used to prove that the double series after the second inequality converges to zero uniformly in points $t^*_{(r)}$ on the lattice. In view of Condition A(i), the single series converges to zero uniformly in $t^*_{(r)}$ as η converges to zero. These results and (8) imply (6).

Step 1.2. Here we show that

$$\sup_{t_{(r)} \in \mathcal{J}_r(c)} |\mathrm{err}_r(t_{(r)}) - \mathrm{err}(t^*_{(r)})| = o(1). \tag{9}$$

The arguments leading to (4) can be used to prove that

$$|\mathrm{err}_r(t_{(r)}) - \mathrm{err}(t^*_{(r)})| \leq \sum_{k=0}^{1} \frac{n_k}{n} \mathrm{pr}_k \Big\{ \big| \pi_k f_k(X \mid t^*_{(r)}) - \pi_{1-k} f_{1-k}(X \mid t^*_{(r)}) \big| \leq |R_6(t_{(r)})| \Big\}, \tag{10}$$

where $\sup_{t_{(r)} \in \mathcal{J}_r(c)} |R_6(t_{(r)})| = o_P(1)$. The latter result implies that, for each $\epsilon > 0$,

$$b_1(t^*_{(r)}) \equiv \mathrm{pr}_k\{|R_6(t^*_{(r)})| > \epsilon\} \to 0 \quad \text{uniformly in} \quad t_{(r)} \in \mathcal{J}_r(c). \tag{11}$$

It follows from Condition A(g) that

$$b_2(t^*_{(r)}) \equiv \mathrm{pr}_k \Big\{ \big| \pi_k f_k(X \mid t^*_{(r)}) - \pi_{1-k} f_{1-k}(X \mid t^*_{(r)}) \big| \leq \epsilon \Big\} \tag{12}$$

uniformly in $t_{(r)} \in \mathcal{J}_r(c)$. Result (10) implies that $|\mathrm{err}_r(t_{(r)}) - \mathrm{err}(t^*_{(r)})| \leq b_1(t^*_{(r)}) + b_2(t^*_{(r)})$, and hence, by (11) and (12), that (9) holds.

Part (i) of Theorem 1 follows from (3) and (9).

Step 2: Part (ii) of the theorem. Part (i) of the theorem implies that $\hat{\mathrm{err}}(t^0_{(r)}) = \mathrm{err}(t^0_{(r)}) + o_P(1)$ and $\hat{\mathrm{err}}(\hat{t}_{(r)}) = \mathrm{err}(\hat{t}_{(r)}) + o_P(1)$. Recall that $t^0_{(r)}$ is contained within a sphere which in turn is contained within $\mathcal{J}_r(c)$. Therefore,

$$\mathrm{err}(\hat{t}_{(r)}) + o_P(1) = \hat{\mathrm{err}}(\hat{t}_{(r)}) \leq \hat{\mathrm{err}}(t^0_{(r)}) = \mathrm{err}(t^0_{(r)}) + o_P(1) \leq \mathrm{err}(\hat{t}_{(r)}) + o_P(1), \tag{13}$$

from which it follows that $\mathrm{err}(\hat{t}_{(r)}) = \mathrm{err}(t^0_{(r)}) + o_P(1)$, i.e. for all $\delta > 0$,

$$\mathrm{pr}\{|\mathrm{err}(\hat{t}_{(r)}) - \mathrm{err}(t^0_{(r)})| > \delta\} \to 0. \tag{14}$$

Condition A(f) implies that, for each $\epsilon > 0$, there exist $\delta > 0$ and $n_0 \geq 1$ such that, for all $n \geq n_0$, $\mathrm{pr}(\|\hat{t}_{(r)} - t^0_{(r)}\| > \epsilon) \leq \mathrm{pr}\{|\mathrm{err}(\hat{t}_{(r)}) - \mathrm{err}(t^0_{(r)})| > \delta\}$, and in conjunction with (14) this implies that $\mathrm{pr}(\|\hat{t}_{(r)} - t^0_{(r)}\| > \epsilon) \to 0$ for all $\epsilon > 0$, which is equivalent to the second part of Theorem 1. □

Proof of Theorem 2. We only prove part (i) since the proof of part (ii) is similar. In the string of identities at (15), the first holds with probability 1 and follows from the definition of $\hat{p}$; the second holds for a random variable $R_7 = R_7(n)$ which, by (13), satisfies $\sup_{r \leq r_0} |R_7| = o_P(1)$; the third holds with probability not less than

$$q_r \equiv \mathrm{pr}\Big[|R_7| < \inf\big\{r \leq r_0 : \mathrm{err}(t^0_{(r+1)}) - (1-\rho)\,\mathrm{err}(t^0_{(r)})\big\}\Big];$$

and the fourth follows from the definition of p:

$$\begin{aligned}
\hat{p} &= \inf\{r \leq r_0 : (1-\rho)\,\mathrm{err}(\hat{t}_{(r)}) \leq \mathrm{err}(\hat{t}^0_{(r+1)})\} \\
&= \inf\{r \leq r_0 : (1-\rho)\,\mathrm{err}(t^0_{(r)}) \leq \mathrm{err}(t^0_{(r+1)}) + R_7\} \\
&= \inf\{r \leq r_0 : (1-\rho)\,\mathrm{err}(t^0_{(r)}) \leq \mathrm{err}(t^0_{(r+1)})\} = p. \qquad (15)
\end{aligned}$$

Now, (A3) and the fact that $R_7 = o_P(1)$ imply that $q_r \to 1$. Hence (15) implies that $\mathrm{pr}(\hat{p} = p) \to 1$ as $n \to \infty$.

Note too that

$$\begin{aligned}
\mathrm{err}^{\mathrm{emp}} &= \frac{n_0}{n}\,\mathrm{pr}_0\big\{J\big(X, \mathcal{D} \mid \hat{t}_{(\hat{p})}\big) = 1\big\} + \frac{n_1}{n}\,\mathrm{pr}_1\big\{J\big(X, \mathcal{D} \mid \hat{t}_{(\hat{p})}\big) = 0\big\} \\
&= \frac{n_0}{n}\,\mathrm{pr}_0\big\{J\big(X, \mathcal{D} \mid \hat{t}_{(p)}\big) = 1\big\} + \frac{n_1}{n}\,\mathrm{pr}_1\big\{J\big(X, \mathcal{D} \mid \hat{t}_{(p)}\big) = 0\big\} + o(1) \\
&= \frac{n_0}{n}\,\mathrm{pr}_0\big\{\pi_0 f_0(X \mid \hat{t}_{(p)}) < \pi_1 f_1(X \mid \hat{t}_{(p)}) + R_8\big\} \\
&\qquad + \frac{n_1}{n}\,\mathrm{pr}_1\big\{\pi_0 f_0(X \mid \hat{t}_{(p)}) > \pi_1 f_1(X \mid \hat{t}_{(p)}) + R_9\big\} + o(1), \qquad (16)
\end{aligned}$$

where, using the uniform convergence of $\hat{f}_k$ to f_k, see (7), R_8 and R_9 denote random variables that equal $o_P(1)$. Remember that the notation $f_k(x \mid t_{(p)})$ refers to the p-dimensional density of $\mathrm{X}(t_{(p)})$ calculated at $\mathrm{x}(t_{(p)})$, when X comes from population k.

The uniform convergence of $\hat{f}_k$ to f_k implies that $\hat{f}_k(x \mid t_{(p)}) = f_k(x \mid t_{(p)}) + o_P(1)$ uniformly in x and $t_{(p)} \in \mathcal{J}_p(c)$, which entails $\hat{f}_k(x \mid \hat{t}_{(p)}) = f_k(x \mid \hat{t}_{(p)}) + o_P(1)$. Hence, by (16),

$$\begin{aligned}
\mathrm{err}^{\mathrm{emp}} &= \frac{n_0}{n}\,\mathrm{pr}_0\{\pi_0 \hat{f}_0(X \mid t^0_{(p)}) < \pi_1 \hat{f}_1(X \mid t^0_{(p)}) + R_{10}\} \\
&\qquad + \frac{n_1}{n}\,\mathrm{pr}_1\{\pi_0 \hat{f}_0(X \mid t^0_{(p)}) > \pi_1 \hat{f}_1(X \mid t^0_{(p)}) + R_{11}\} + o(1) \\
&= \frac{n_0}{n}\,\mathrm{pr}_0\{\pi_0 \hat{f}_0(X \mid t^0_{(p)}) < \pi_1 \hat{f}_1(X \mid t^0_{(p)})\} \\
&\qquad + \frac{n_1}{n}\,\mathrm{pr}_1\{\pi_0 \hat{f}_0(X \mid t^0_{(p)}) < \pi_1 \hat{f}_1(X \mid t^0_{(p)})\} + o(1) = \mathrm{err}(t^0_{(p)}) + o(1),
\end{aligned}$$

where the second last equality is obtained using calculations similar to those in the proof of Theorem 1. □