ON ERROR-RATE ESTIMATION IN NONPARAMETRIC CLASSIFICATION

Anil Ghosh[1] and Peter Hall[1,2]

ABSTRACT. There is a substantial literature on the estimation of error rate, or risk, for nonparametric classifiers. Error-rate estimation has at least two purposes: accurately describing the error rate, and estimating the tuning parameters that permit the error rate to be mininised. In the light of work on related problems in nonparametric statistics, it is attractive to argue that both problems admit the same solution. Indeed, methods for optimising the point-estimation performance of nonparametric curve estimators often start from an accurate estimator of error. However, we argue in this paper that accurate estimators of error rate in classification generally give poor results when used to choose tuning parameters; and vice versa. Concise theory is used to illustrate this point in the case of cross-validation (which gives very accurate estimators of error rate, but poor estimators of tuning parameters) and the smoothed bootstrap (where error-rate estimation is poor but tuning-parameter approximations are particularly good). The theory is readily extended to other methods, for example to the .632+ bootstrap approach, which gives good estimators of error rate but poor estimators of tuning parameters. Reasons for the apparent contradiction are given, and numerical results are used to point to the practical implications of the theory.

KEYWORDS. Bayes risk, bootstrap, cross-validation, classification error, discrimination, error rate, kernel methods, nonparametric density estimation, risk.

SHORT TITLE. Nonparametric classification.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Department of Mathematics and Statistics, The University of Melbourne, Parkville, Vic 3052, Australia

1. INTRODUCTION

Cross-validation is a widely used technique for estimating the risk, or error rate, of classification procedures. It often gives estimators which are close to unbiased, and which have good mean squared error properties. In fact, cross-validation is frequently viewed as the method to beat when alternative approaches to risk estimation are suggested; see, for example, Efron (1983) and Efron and Tibshirani (1997).

One might expect that good performance in estimating risk would be accompanied by good performance in determining the values of tuning parameters that minimise risk. Indeed, in a number of related model-selection problems, computing a good estimator of error is the first step in approximating the values of the parameters that minimise error. Early work of this type includes that of Hall (1983), Bowman (1984), Stone (1984) and Faraway and Jhun (1990).

However, in important ways the problem of risk estimation in classification is significantly different from a number of apparently similar problems in nonparametric statistics. A major reason is that the tuning parameters used to construct classifiers may influence performance only in a relatively small number of places, for example the places where population densities cross. Therefore, the impact that the parameters have on risk can be relatively minor. As a result, in classification problems it is possible to construct a particularly accurate estimator of risk which is of very little value for estimating tuning parameters; cross-validation turns out to be of this type. The contrary case also arises — empirical risk-based methods for estimating turning parameters may perform that task very well, but give poor estimators of risk.

It is unsurprising that this problem is not well understood. Indeed, it is somewhat contradictory to argue that one should not seek an accurate estimator of risk when attempting to minimise that quantity empirically. However, a consequence of not fully understanding the problem is that methods such as cross-validation, and its jackknife or bootstrap competitors, which are designed to minimise risk, are in practice pressed into service to select tuning parameters. This can be inappropriate.

In the present paper we shall point to the shortcomings of cross-validation for estimating tuning parameters for classifiers, and also to the advantages of other approaches that give accurate estimators of tuning parameters but poor estimators of risk. In order to make our discussion and technical arguments transparent, we shall treat a relatively simple, univariate problem, where standard kernel estimators are used as the basis for classifiers. However, similar results can be derived in

multivariate settings, and also when methods other than kernel estimators are used for classification. In the kernel case, the tuning parameters referred to above are bandwidths.

In this paper we shall derive theoretical results that address the following points. (a) The cross-validation estimator of risk is root-n consistent, and in fact is asymptotically equivalent to a nonparametric maximum likelihood estimator. (b) Notwithstanding property (a), the part of the cross-validation estimator that depends on the tuning parameters is very highly stochastically variable, so much so that it has an unboundedly large number of local extrema which bear no important asymptotic relationship to the parameters that actually minimise risk. (c) In marked contrast to (a), a smoothed bootstrap estimator of risk is very highly biased, and in consequence can have poor convergence rates relative to the cross-validation approach. (d) Despite the drawbacks noted in (c), the smoothed bootstrap method produces accurate estimators of the parameters that minimise risk. (e) The smoothed bootstrap method is particularly robust against inappropriate choice of smoothing parameters, and in fact those quantities can be selected within a very broad range without appreciably influencing performance.

Property (e) will be reinforced by our numerical work in section 3, which will also introduce an adaptive, empirical approach to smoothing the bootstrap. Property (b) has been discussed by Hall and Kang (2006), but without a concise mathematical account of the issues involved. The erratic way in which the cross-validation criterion varies with tuning parameters is well known.

More generally, the theoretical results given in the present paper can be augmented by others, which show that some of the problems that afflict cross-validation can be reduced by applying Breiman's (1996) bagging technique to dampen down the effects of excessive variability. However, it seems difficult to achieve the good performance of the smoothed bootstrap approach, and for that reason we omit from this paper a theoretical treatment of bagged cross-validation.

The fact that bagging does not redress all the problems associated with cross-validation is highlighted in our numerical work, where we treat three additional approaches: bagged cross-validation, and the bootstrap methods suggested by Efron (1983) and Efron and Tibshirani (1997), respectively. In terms of their performance at estimating tuning parameters, the first and second of these techniques lie between cross-validation (at the lower-performance end of the scale) and the smoothed bootstrap (at the upper end). The method of Efron and Tibshirani

(1997) gives particularly poor estimators of tuning parameters.

There is an especially large literature on nonparametric classification. Kernel-based approaches date from work of Fix and Hodges (1951). More generally, a large variety of classification methodologies has been developed based on empirical forms of the Bayes classifier. Relatively recent contributions include those of Chanda and Ruymgaart (1989), Krzyżak (1991), Lapko (1993), Pawlak (1993), Lugosi and Pawlak (1994), Devroye, Györfi and Lugosi (1996), Lugosi and Nobel (1996), Ancukiewicz (1998), Yang (1999a,b), Mammen and Tsybakov (1999), Steele and Patterson (2000) and Lin (2001).

2. MAIN RESULTS

2.1. Error rates and their estimators. Let F and G denote distributions with respective densities f and g, and let $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ be datasets drawn respectively from F and G. We consider ways of classifying a new data value, x say, to either the F or the G population.

Assume that the distributions F and G have respective prior probabilities p and $1 - p$. Define

$$\Delta = p\, f - (1 - p)\, g\,. \tag{2.1}$$

The Bayes classifier, $\mathcal{A}_0$ say, allocates x to F or G according as $\Delta(x)$ is positive or negative, respectively. The corresponding error rate, or risk, for classification of data on a compact interval $\mathcal{I}$, is

$$\mathrm{err}_{\mathcal{A}_0} = p \int_{\mathcal{I}} I\{\Delta(x) < 0\}\, f(x)\, dx + (1 - p) \int_{\mathcal{I}} I\{\Delta(x) > 0\}\, g(x)\, dx\,.$$

A general class of classifiers can be constructed by replacing Δ by an estimator, $\widehat{\Delta}$. As a prelude to defining $\widehat{\Delta}$ we introduce density estimators $\hat{f}$ and $\hat{g}$, and their leave-one-out versions $\hat{f}_{-i}$ and $\hat{g}_{-i}$. Let K be a nonnegative kernel and h_1 and h_2 be bandwidths, and put $m_1 = m - 1$, $n_1 = n - 1$,

$$\begin{aligned} \hat{f}(x) &= \frac{1}{m h_1} \sum_{i=1}^{m} K\biggl(\frac{x - X_i}{h_1}\biggr)\,, \qquad & \hat{g}(x) &= \frac{1}{n h_2} \sum_{i=1}^{n} K\biggl(\frac{x - Y_j}{h_2}\biggr)\,, \\ \hat{f}_{-i}(x) &= \frac{1}{m_1 h_1} \sum_{j=1}^{m}{}^{(i)}\, K\biggl(\frac{x - X_i}{h_1}\biggr)\,, & \hat{g}_{-i}(x) &= \frac{1}{n_1 h_2} \sum_{j=1}^{n}{}^{(i)}\, K\biggl(\frac{x - Y_j}{h_2}\biggr)\,, \end{aligned} \tag{2.2}$$

where $\sum_j^{(i)}$ denotes summation over indices j not equal to i. Define

$$\widehat{\Delta} = p\, \hat{f} - (1 - p)\, \hat{g}\,.$$

The empirical classifier $\mathcal{A}_1$, which assigns x to distribution F if $\widehat{\Delta}(x) > 0$, and to G otherwise, has the following empirical risk:

$$\text{emperr}_{\mathcal{A}_1}(h_1, h_2) = p \int_{\mathcal{I}} I\big\{\widehat{\Delta}(x) < 0\big\}\, f(x)\, dx + (1-p) \int_{\mathcal{I}} I\big\{\widehat{\Delta}(x) > 0\big\}\, g(x)\, dx\,.$$

Its average value over all possible datasets $\mathcal{X}$ and $\mathcal{Y}$, i.e. its expected value, is

$$\text{err}_{\mathcal{A}_1}(h_1, h_2) = p \int_{\mathcal{I}} P\big\{\widehat{\Delta}(x) < 0\big\}\, f(x)\, dx + (1-p) \int_{\mathcal{I}} P\big\{\widehat{\Delta}(x) > 0\big\}\, g(x)\, dx\,.$$

The cross-validation estimator of $\text{err}_{\mathcal{A}_1}$ is

$$\begin{aligned}\text{CV}(h_1, h_2) = \frac{p}{m} \sum_{i=1}^{m} & I\big\{\widehat{\Delta}_{f,-i}(X_i) < 0, X_i \in \mathcal{I}\big\} \\ & + \frac{1-p}{n} \sum_{i=1}^{n} I\big\{\widehat{\Delta}_{g,-i}(Y_i) > 0, Y_i \in \mathcal{I}\big\}\,, \qquad (2.3)\end{aligned}$$

where $\widehat{\Delta}_{f,-i} = p\, \hat{f}_{-i} - (1-p)\, \hat{g}$ and $\widehat{\Delta}_{g,-i} = p\, \hat{f} - (1-p)\, \hat{g}_{-i}$.

An alternative way of estimating risk is to resample from smoothed versions of the empirical distributions of X and Y, conditional on the data in $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, using bootstrap methods. Specifically, let $\tilde{f}$ and $\tilde{g}$ denote the versions of $\hat{f}$ and $\hat{g}$, respectively, at (2.2), constructed using bandwidths h_3 and h_4 in place of h_1 and h_2. (Potentially, the bandwidths employed at this point can be quite different from those used to construct the classifier.) Let $\mathcal{X}^* = \{X_1^*, \ldots, X_m^*\}$ and $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_n^*\}$ denote datasets drawn by sampling randomly, conditional on $\mathcal{Z}$, from the distributions with respective densities $\tilde{f}$ and $\tilde{g}$. Construct the versions $\hat{f}^*$ and $\hat{g}^*$ of $\hat{f}$ and $\hat{g}$ from these resamples, on this occasion using the original bandwidths h_1 and h_2:

$$\hat{f}^*(x) = \frac{1}{mh_1} \sum_{i=1}^{m} K\left(\frac{x - X_i^*}{h_1}\right), \qquad \hat{g}^*(x) = \frac{1}{nh_2} \sum_{i=1}^{n} K\left(\frac{x - Y_j^*}{h_2}\right).$$

Put $\widehat{\Delta}^* = p\, \hat{f}^* - (1-p)\, \hat{g}^*$. The bootstrap classifier consists of assigning a new data value x to F if $\widehat{\Delta}^*(x) > 0$, and assigning it to G otherwise.

The long-run error rate of this classifier, conditional on the data $\mathcal{Z}$, is given by

$$\begin{aligned}&\widehat{\text{err}}_{\mathcal{A}_1}(h_1, h_2) \\ &\quad = p \int_{\mathcal{I}} P\big\{\widehat{\Delta}^*(x) < 0 \,\big|\, \mathcal{Z}\big\}\, \tilde{f}(x)\, dx + (1-p) \int_{\mathcal{I}} P\big\{\widehat{\Delta}^*(x) \,\big|\, \mathcal{Z} > 0\big\}\, \tilde{g}(x)\, dx\,.\end{aligned}$$

In particular, $\widehat{\text{err}}_{\mathcal{A}_1}$ is an approximation to $\text{err}_{\mathcal{A}_1}$ alternative to the cross-validation criterion, CV.

2.2. Summary of properties of CV, emperr *and* $\widehat{\text{err}}$ *as approximations to risk.* It is known (see e.g. Hall and Kang 2005) that the bandwidths that are optimal in the sense of minimising the risk $\text{err}_{\mathcal{A}_1}$ of the empirical classification rule $\mathcal{A}_1$, are generally of the same size as $h = n^{-1/5}$. Therefore we shall assess risks when the bandwidths are on this scale.

In this context, the "regret" of the rule $\mathcal{A}_1$ is of size h^4. That is, when both h_1 and h_2 are of size $n^{-1/5}$, the difference between the risk $\text{err}_{\mathcal{A}_1}$ for $\mathcal{A}_1$, and the risk $\text{err}_{\mathcal{A}_0}$ for the Bayes classifier, is asymptotic to a constant multiple of h^4. (See e.g. Hall and Kang (2005), and also (2.14) below.) The regret captures all of the influence that bandwidth choice has on risk. Moreover, the empirical risk, $\text{emperr}_{\mathcal{A}_1}$, also equals the Bayes risk plus a term of order h^4. However, in the case of $\text{emperr}_{\mathcal{A}_1}$ the h^4 term includes non-negligible stochastic fluctuations, which depend substantially on h_1 and h_2 but in a manner that does not reflect the dependence of $\text{err}_{\mathcal{A}_1}$ on those bandwidths.

We shall show that the cross-validation estimator of risk equals a term which does not depend on either of the bandwidths h_1 and h_2, plus a highly stochastically volatile quantity which is of size $h^{7/2}$, and so is an order of magnitude larger than h^4. More particularly, the fluctuations of the cross-validation criterion CV, as a function of the bandwidths h_1 and h_2, are largely unrelated to the ways in which either the risk $\text{err}_{\mathcal{A}_1}$, or its empirical form $\text{emperr}_{\mathcal{A}_1}$, are influenced by the bandwidths. Therefore, minimising the cross-validation criterion does not correspond, in an asymptotic or in another meaningful sense, to minimising either the true or the empirical risk.

However, we shall note that $\text{CV}(h_1, h_2)$ is very close to being an unbiased estimator of $\text{err}_{\mathcal{A}_1}(h_1, h_2)$, indeed so close that it captures all the main effects of bandwidth choice on risk. Therefore, the difficulties that afflict cross-validation arise from stochastic variability, not systematic error.

The strengths and weaknesses of $\widehat{\text{err}}_{\mathcal{A}_1}$, as an alternative to CV, are diametically opposite to those of CV. In particular, $\widehat{\text{err}}_{\mathcal{A}_1}$ suffers from substantial bias as an estimator $\text{err}_{\mathcal{A}_1}$, but the stochastic variability of that portion of $\widehat{\text{err}}_{\mathcal{A}_1}$ that captures the main effects of bandwidth is particularly low. As a result, $\widehat{\text{err}}_{\mathcal{A}_1}$ can be used effectively to choose bandwidths for density-based classifiers.

2.3. Details of properties of cross-validation. We shall assume that:

$$\begin{aligned}&K \text{ is symmetric, compactly supported, integrates to 1 and has two bounded derivatives; the function } \Delta, \text{ defined at (2.1), vanishes in } \mathcal{I} \text{ only at } r \text{ isolated points, say } y_1, \ldots, y_r, \text{ in the interior of } \mathcal{I}, \text{ and at each point } y_i,\ \Delta'(y_i) \neq 0; \\ &f \text{ and } g \text{ are continuous on the real line, and have two Hölder-continuous derivatives in an open neighbourhood of } y_i \text{ for } 1 \leq i \leq r; \ m \text{ and } n \text{ increase together, and } n/m \to \rho, \text{ where } 0 < \rho < \infty.\end{aligned} \tag{2.4}$$

Put $h = n^{-1/5}$. Given $B > 1$, let $\mathcal{H} = \mathcal{H}(B)$ denote the set of values of hu for which $B^{-1} \leq u \leq B$. We shall take both the bandwidths h_1 and h_2, used to construct the estimators at (2.2), to be in $\mathcal{H}$.

For $1 \leq i \leq r$, let $\mathcal{W}_{Xi}$, $\mathcal{W}_{Yi}$, $\mathcal{G}_{Xi}$ and $\mathcal{G}_{Yi}$ denote independent stochastic processes, with $\mathcal{W}_{Xi}$ and $\mathcal{W}_{Yi}$ being standard Wiener processes, and $\mathcal{G}_{Xi}$ and $\mathcal{G}_{Yi}$ Gaussian processes having zero means and covariance given by, for both $Z = X$ and $Z = Y$,

$$\operatorname{cov}\bigl\{\mathcal{G}_{Zi}\bigl(t^{(1)}\bigr)\,,\,\mathcal{G}_{Zi}\bigl(t^{(2)}\bigr)\bigr\} = \int K\bigl(s\,t^{(1)}\bigr)\,K\bigl(s\,t^{(2)}\bigr)\,ds\,. \tag{2.5}$$

Put $\kappa = \int K^2$ and $\kappa_2 = \int x^2\,K(x)\,dx$. Given $0 < \rho < \infty$ and $u_1, u_2 > 0$, let $u = (u_1, u_2)$,

$$d(y\,|\,u) = \tfrac{1}{2}\,\kappa_2\,\bigl\{p\,u_1^2\,f''(y) - (1-p)\,u_2^2\,g''(y)\bigr\}\,, \tag{2.6}$$

$$\mathcal{V}_i(u) = \frac{1}{\Delta'(y_i)}\,\Bigl[p\,\{\rho\,f(y_i)\}^{1/2}\,\mathcal{G}_{Xi}(u_1) + (1-p)\,g(y_i)^{1/2}\,\mathcal{G}_{Yi}(u_2) + d(y_i\,|\,u)\Bigr]\,, \tag{2.7}$$

$$\begin{aligned}T(u) &= \sum_{i=1}^r\,\Bigl[p\,\{\rho\,f(y_i)\}^{1/2}\,\mathcal{W}_{Xi}\{\mathcal{V}_i(u)\} + (1-p)\,g(y_i)^{1/2}\,\mathcal{W}_{Yi}\{\mathcal{V}_i(u)\}\Bigr]\,, \\ \tau(u) &= \tfrac{1}{2}\,\kappa\sum_{i=1}^r\,|\Delta'(y_i)|^{-1}\,\bigl\{p^2\,\rho\,f(y_i)\,u_1^{-1} + (1-p)^2\,g(y_i)\,u_2^{-1}\bigr\} \\ &\qquad + \tfrac{1}{2}\,\sum_{i=1}^r\,|\Delta'(y_i)|^{-1}\,d(y_i\,|\,u)^2\,.\end{aligned} \tag{2.8}$$

If $\mathcal{I} = [a, b]$, define $y_0 = a$ and $y_{r+1} = b$, and for $1 \leq i \leq r+1$, let L_i denote the number of indices j for which $y_{i-1} < X_j < y_i$ if $\Delta < 0$ on (y_{i-1}, y_i), or the number of j for which $y_{i-1} < Y_j < y_i$ if $\Delta > 0$ on (y_{i-1}, y_i). Put $p_i = p/m$ in the first of these cases, and $p_i = (1-p)/n$ in the second.

Theorem 2.1. *Assume conditions* (2.4). *Then the stochastic processes $\mathcal{W}_{Xi}$, $\mathcal{W}_{Yi}$ and $\mathcal{V}_i$ can be constructed, depending in each instance on n, such that, with $h_j = hu_j$ for $j = 1$ and 2,*

$$\mathrm{CV}(h_1, h_2) = \sum_{i=1}^{r+1}\,p_i\,L_i + h^{7/2}\,T(u) + o_p\bigl(h^{7/2}\bigr)\,, \tag{2.9}$$

where the remainder is of the stated order uniformly in $B^{-1} \le u_1, u_2 \le B$, *with* $B > 1$. *Furthermore,*

$$E\{\mathrm{CV}(h_1, h_2)\} = \mathrm{err}_{\mathcal{A}_1} + o\big(h^4\big) \tag{2.10}$$

$$= \mathrm{err}_{\mathcal{A}_0} + \tfrac{1}{2} \sum_{i=1}^{r} |\Delta'(y_i)|^{-1} E\big\{p\,\hat{f}(y_i) - (1-p)\,\hat{g}(y_i)\big\}^2 + o\big(h^4\big) \tag{2.11}$$

$$= \mathrm{err}_{\mathcal{A}_0} + h^4\,\tau(u) + o\big(h^4\big)\,, \tag{2.12}$$

again uniformly in $B^{-1} \le u_1, u_2 \le B$.

The two main terms on the right-hand side of (2.9), i.e. the series $\sum_i p_i\,L_i$ and the subsequent term $h^{7/2}\,T(u)$, represent a division of $\mathrm{CV}(h_1, h_2)$ into parts that represent, respectively, the dominant part of CV that does not depend on the bandwidths h_1 and h_2, and the dominant part that is influenced by those bandwidths. In particular, the series $\sum_i p_i\,L_i$ does not depend on h_1 and h_2. Its expected value equals the risk of the Bayes rule for classification on the interval $\mathcal{I}$:

$$E\bigg(\sum_{i=1}^{r+1} p_i\,L_i\bigg) = \mathrm{err}_{\mathcal{A}_0}\,.$$

The variance of $\sum_i p_i\,L_i$ is of order n^{-1}.

If we were given the values of $y_1, \ldots, y_r$, and told also the signs of Δ on the intervals between adjacent values of y_i, then the nonparametric maximum likelihood estimator of $\mathrm{err}_{\mathcal{A}_0}$ would be exactly $\sum_i p_i\,L_i$. In particular, this series has minimum variance among all unbiased estimators of $\mathrm{err}_{\mathcal{A}_0}$, and its convergence rate, $n^{-1/2}$, cannot be improved upon. Therefore, the series $\sum_i p_i\,L_i$, which represents the dominant part of CV and converges to its expected value at the slower of the rates for the two respective terms in (2.9), cannot be made significantly more accurate as an approximation to the Bayes-rule risk $\mathrm{err}_{\mathcal{A}_0}$.

However, the term $\sum_i p_i\,L_i$ does not provide any information about the effect of bandwidth on classification performance in neighbourhoods of the points $y_1, \ldots, y_r$. Of course, that information is crucial to understanding how properties of the classifier are influenced by its contruction. We have to pass to the second term, $h^{7/2}\,T(u)$, on the right-hand side of (2.9), in order to obtain any information about how h_1 and h_2 influence $\mathrm{CV}(h_1, h_2)$.

Revealingly, the second term varies stochastically in a very erratic manner. Indeed, since the Wiener processes $\mathcal{W}_{Xi}$ and $\mathcal{W}_{Yi}$ have fractal sample paths then,

with probability 1, $T(u_1, u_2)$ has an infinite number of local minima, as a function of u_1 and u_2, in any rectangle. Indicative of this high degree of volatility, a graph of a realisation of the function $v = T(u_1, u_2)$, as a function of u_1 and u_2, is, with probability 1, a surface of fractal dimension exceeding 2.

These difficulties persist even if we take $u_1 = u_2 = t$, say. (That choice reflects a common practice of using the same bandwidth to construct density estimators from either dataset.) Of course, in practice the cross-validation criterion $\mathrm{CV}(ht, ht)$ has only a finite number of local maxima in any nondegenerate interval, but the fact that the stochastic approximant $T(t, t)$ has an infinite number of local maxima there implies that the number of local minima of $\mathrm{CV}(ht, ht)$ in the interval increases without unbound as sample size diverges.

These results provide a theoretical explanation of the observed high degree of stochastic variability of the cross-validation criterion, in terms of the way it describes the effect of bandwidth choice on risk. The results also indicate why choosing the bandwidths to minimise CV is frought with practical difficulty. When using cross-validation with real data it is found that the criterion has many local minima, few of which seem more apropriate than the others; and that this problem becomes more, rather than less, pronounced as sample size increases. Both these properties are implied by the theoretical results discussed in the two previous paragraphs, and so the theory provides insight into practice.

On the other hand, results (2.10)–(2.12) assert that the cross-validation criterion is close to being an unbiased approximation to the risk of the empirical rule based on $\widehat{\Delta}$.

2.4. Properties of empirical risk. Here we state and discuss a version of Theorem 2.1 for $\mathrm{emperr}_{\mathcal{A}_1}$, rather than for the cross-validation approximation to the risk. The empirical risk is not computable in practice, but from some viewpoints one would not expect the cross-validation approximation to be markedly inferior to the empirical risk, at least in terms of the way it reflects properties of the bandwidths. In fact, it is substantially inferior.

Recall that $u = (u_1, u_2)$, and define $\hat{\tau} = \tau + \hat{\tau}_1$, where τ is as at (2.8) and

$$\hat{\tau}_1(u) = \sum_{i=1}^{r} \Delta'(y_i) \int \Big[I\{v < \mathcal{V}_i(u)\} - P\{v < \mathcal{V}_i(u)\} \Big] \, dv \,.$$

Theorem 2.2. *Assume conditions* (2.4). *Then, with* $\mathcal{V}_i$ *as in Theorem 2.1, and*

with $h_j = hu_j$ for $j = 1$ and 2,

$$\begin{aligned}\mathrm{emperr}_{\mathcal{A}_1}(h_1, h_2) &= \mathrm{err}_{\mathcal{A}_1}(h_1, h_2) + h^4\,\hat\tau_1(u) + o_p\big(h^4\big) \\ &= \mathrm{err}_{\mathcal{A}_0} + h^4\,\hat\tau(u) + o_p\big(h^4\big)\,, \end{aligned} \tag{2.13}$$

where the remainders are of the stated orders uniformly in $B^{-1} \le u_1, u_2 \le B$, for any $B > 1$.

A formula for the error of the classification rule $\mathcal{A}_1$ is obtainable directly from (2.10) and (2.12):

$$\mathrm{err}_{\mathcal{A}_1} = \mathrm{err}_{\mathcal{A}_0} + h^4\,\tau(u) + o\big(h^4\big)\,, \tag{2.14}$$

where the positive function τ is given at (2.8). In particular, (2.14) implies that the regret is asymptotic to a constant multiple of h^4. Result (2.13) shows that the difference between the empirical risk and the actual risk is of the same size as, but not asymptotically equal to, the regret. The former, multiplied by h^{-4}, converges in distribution to a nondegenerate random variable which can take both positive and negative values, whereas the regret, multiplied by h^{-4}, converges to a positive constant.

More importantly, a comparison of (2.9), (2.13) and (2.14) shows that there is no useful connection between the parts of $\mathrm{CV}(h_1, h_2)$ that depend on h_1 and h_2, and the corresponding parts of either the risk $\mathrm{err}_{\mathcal{A}_1}$ or its empirical version $\mathrm{emperr}_{\mathcal{A}_1}$. In particular, the term $h^{7/2}\,T(u)$ in (2.9) is an order of magnitude larger than both $h^4\,\tau(u)$ and $h^4\,\hat\tau(u)$, on the right-hand sides of (2.14) and (2.13) respectively. Moreover, the fluctuations of $T(u)$, as a function of u, bear no relationship to those of $\tau(u)$ or $\hat\tau(u)$. Therefore, cross-validation cannot be used effectively to choose the bandwidths that minimise either $\mathrm{err}_{\mathcal{A}_1}$ or $\mathrm{emperr}_{\mathcal{A}_1}$.

2.5. Properties of bootstrap estimator of risk. In the light of what we have learned in earlier sections, the properties of $\widehat{\mathrm{err}}_{\mathcal{A}_1}(h_1, h_2)$ are relatively transparent. In particular, provided the bandwidths h_3 and h_4 (used to construct the density estimators $\tilde f$ and $\tilde g$) are sufficiently large to ensure that $\tilde f''$ and $\tilde g''$ are consistent for f'' and g'', respectively, the bootstrap analogue of (2.14) holds:

$$\widehat{\mathrm{err}}_{\mathcal{A}_1}(h_1, h_2) = \widehat{\mathrm{err}}_{\mathcal{A}_0} + h^4\,\tau(u) + o_p\big(h^4\big)\,, \tag{2.15}$$

uniformly in $B^{-1} \le u_1, u_2 \le B$. In (2.15), $\widehat{\mathrm{err}}_{\mathcal{A}_0}$ denotes the estimator of $\mathrm{err}_{\mathcal{A}_0}$ that is obtained on replacing, in the definition of $\mathrm{err}_{\mathcal{A}_0}$, the unknown densities f and g by their estimators $\tilde f$ and $\tilde g$:

$$\widehat{\mathrm{err}}_{\mathcal{A}_0} = p \int_{\mathcal{I}} I\big\{\widetilde\Delta(x) < 0\big\}\,\tilde f(x)\,dx + (1-p) \int_{\mathcal{I}} I\big\{\widetilde\Delta(x) > 0\big\}\,\tilde g(x)\,dx\,,$$

where $\widetilde{\Delta} = p\,\tilde{f} - (1-p)\,\tilde{g}$. Note particularly that $\widehat{\mathrm{err}}_{\mathcal{A}_0}$ does not depend on the bandwidths h_1 and h_2.

The quantity $\widehat{\mathrm{err}}_{\mathcal{A}_0}$ will generally not be a good estimator of $\mathrm{err}_{\mathcal{A}_0}$. In particular, the relatively large values needed for the bandwidths h_3 and h_4 will ensure that $\widehat{\mathrm{err}}_{\mathcal{A}_0}$ suffers from significant bias, although (e.g. under the conditions of Theorem 2.3 below) it will be consistent. However, this inaccuracy is not necessarily a problem if our aim is determine, from $\widehat{\mathrm{err}}_{\mathcal{A}_1}$, the influence that h_1 and h_2 have on the true risk, $\mathrm{err}_{\mathcal{A}_1}$. Since $\widehat{\mathrm{err}}_{\mathcal{A}_0}$ does not depend on h_1 and h_2 then the main effect of the influence of those quantities is expressed through the term $h^4\,\tau(u)$ on the right-hand side of (2.15), and so is exactly the same as main effect of the influence of h_1 and h_2 on $\mathrm{err}_{\mathcal{A}_1}$; see (2.14). Hence, we can use $\widehat{\mathrm{err}}_{\mathcal{A}_1}$ effectively to choose the bandwidth that minimises risk.

Theorem 2.3. *Assume conditions* (2.4)*, and that the bandwidths h_3 and h_4 both satisfy $n^{(1/5)-\epsilon}h_j \to \infty$ and $n^{\epsilon}h_j \to 0$ for some $\epsilon > 0$. Then* (2.15) *holds, uniformly in $B^{-1} \le u_1, u_2 \le B$.*

A proof of Theorem 2.3 is similar to that of Theorem 2.1, and so will not be given.

3. NUMERICAL PROPERTIES

In this section we shall report the results of a simulation study addressing numerical properties of risk estimators based on cross-validation and the bootstrap. We know from our theoretical work that having an estimator of risk that is good for estimating bandwidth, is not necessarily the same as having a good estimator of risk itself. For example, we showed in section 2 that the bootstrap gives an estimator of risk that is seriously biased, relative to the estimator produced by cross-validation; and that the bootstrap approach nevertheless gives better bandwidth estimators. Both these results are reflected starkly in numerical experiments, although we shall report here only the results about bandwidth choice.

Likewise, Efron's (1983) bootstrap method is known not to be a good estimator of risk, since it usually leads to overestimation (see e.g., Efron and Tibshirani, 1997), but here we show that it nevertheless performs reasonably well as a bandwidth selector. On the other hand, Efron and Tibshirani's (1997) .632+ bootstrap method gives very good estimators of risk, as that paper shows, but it does not perform well when used to select bandwidth, as we demonstrate below.

For the most part we consider the case where the distributions with densities

f and g are equal-probability mixtures of two univariate normal distributions, with means 0 and 2 (in the case of f) and 1 and 3 (for g), each component having variance 0.25. The context where the Normal distributions are both replaced by lognormal distributions, or by Cauchy distributions, will also be discussed. We take $p = \frac{1}{2}$ and $m = n$, and $h_2 = sh_1$, where s denotes the ratio of a measure of the scale of f to that for g. This approach is often used in practice.

The first panel of Figure 3.1 depicts 100 plots of $\mathrm{CV}(h_1, sh_1)$ as a function of h_1, when s is taken equal to 1. Since the two population distributions have the same variance then this choice of s is reasonable. We shall discuss shortly the case where s is estimated from data.

Each curve in Figure 3.1 is computed for a different pair $(\mathcal{X}, \mathcal{Y})$ of random samples, of sizes $m = n = 100$. The second panel of the figure shows, for the same 100 sample pairs, plots of the bootstrap alternative to the cross-validation criterion, $\widehat{\mathrm{err}}_{\mathcal{A}_1}(h_1, sh_1)$. We used $h_3 = h_4 = 0.3$ when computing the bootstrap density estimators $\hat{f}^*$ and $\hat{g}^*$.

The erratic nature of the graphs in the first panel of Figure 3.1 reflects the high degree of variability of cross-validation, demonstrated theoretically in section 2. This suggests that cross-validation has substantial difficulty, relative to the bootstrap, approximating the optimal bandwidth. That is confirmed by Figure 3.2, which gives a histogram estimator of the distribution of the bandwidths that minimise $\mathrm{CV}(h_1, sh_1)$ (in the left-hand panel of Figure 3.2) or $\widehat{\mathrm{err}}_{\mathcal{A}_1}(h_1, sh_1)$ (in the right-hand panel). In this setting the theoretically optimal bandwidth, in the sense of minimising risk under the constraint $h_2 = sh_1$, can be shown to be $h_1 = 0.26$, which value is indicated by a small black triangle on the horizontal axes. The bootstrap bandwidth estimator is close to being unbiased, and has low stochastic variability, whereas the cross-validation estimator is skewed to the right and is very highly variable. Results for different sample sizes, and for other densities f and g, are similar.

In order to improve the performance of cross-validation we used the bootstrap aggregation, or bagging, technique suggested by Breiman (1996). From each sample, a proportion α (where $0 < \alpha < 1$) of data was resampled, without replacement, to form a new subsample. Cross-validation was applied to this subsample to estimate the risk function. This step was repeated many times, and an overall estimator of risk was obtained by averaging. The first panel of Figure 3.3 shows that the resulting, bagged version of $\mathrm{CV}(h_1, sh_1)$ has substantially lower stochastic variation

than its unbagged counterpart.

Of course, when using the bagged form of $\mathrm{CV}(h_1, sh_1)$ to select bandwidth, we need to correct for the fact that we reduced sample size by the factor α. As discussed in section 2, it is known that the optimal bandwidth is of size $n^{-1/5}$, and so an appropriate correction is readily obtained by taking the bandwidth that minimises the bagged form of $\mathrm{CV}(h_1, sh_1)$, and reducing it by multiplying by the factor $\alpha^{1/5}$.

As an aid to determining the appropriate value of α we experimented with different training-sample sizes ($m = n = 50$, 100, 150 and 200) and different values of α. In each case, we generated 100 different training samples and computed the true risk function corresponding to the selected bandwidth. Average values of these risk functions are reported in Figure 3.4, for different values of α and n. It can be seen that, for $n \geq 100$, the method is largely unaffected by different choices of α, although values in the range $0.2 \leq \alpha \leq 0.4$ are mildly preferable. Similar results are obtained for other density pairs (f, g). Therefore we take $\alpha = 0.3$ in the work below.

Analogous experiments were conducted to determine appropriate formulae for the plug-in bandwidths h_3 and h_4 used in the bootstrap algorithm. Echoing the very wide theoretical range permitted in Theorem 2.3, and reflecting our experience when determining the amount of bagging that should be applied, the effect of choice of smoothing parameter was found to be particularly small for $n \geq 100$. Indeed, choice of h_3 and h_4 does not seem to be a significant issue. Our numerical experiments suggest that if the training-sample sizes are approximately equal then the choices $h_3 = n^{0.05}\,\hat{h}_1$ and $h_4 = n^{0.05}\,s\,\hat{h}_1$ are appropriate in the bootstrap stage, where $\hat{h}_1$ is the optimum bandwidth estimated by the bagged version of cross-validation.

Figure 3.5 depicts the relative performances of two of the bandwidth selectors discussed above: bagged cross-validation with sampling fraction equal to 0.3, and the bootstrap with empirical choice of h_3 and h_4. (Figure 3.2 gave the analogous histogram in the case of cross-validation.) For comparison we also include the bootstrap method suggested by Efron (1983), and the .632+ bootstrap method proposed by Efron and Tibshirani (1997).

Like standard cross-validation based on $\mathrm{CV}(h_1, sh_1)$, the .632+ bootstrap does a good job estimating risk for its own sake, and in particular produces estimators that are significantly less biased than those given by the bootstrap criterion $\widehat{\mathrm{err}}_{\mathcal{A}_1}(h_1, sh_1)$. However, also like cross-validation, it has poor performance

when used to estimate bandwidth. Likewise, Efron's (1983) method is superior to $\widehat{\text{err}}_{\mathcal{A}_1}$ at estimating risk; the fact that it is inferior to $\widehat{\text{err}}_{\mathcal{A}_1}$ when used to choose bandwidth is not a contradiction.

In practice one would use an estimator, $\hat{s}$ say, of s, for example the ratio of standard deviations or of the interquartile ranges. In the normal-mixture case, results obtained for either of these approaches were virtually identical to that when $s = 1$. In particular, Figures 3.1–3.5 were almost unchanged. The results reported for the remainder of this section will be for the case where s was replaced by the ratio of interquartile ranges.

Complementing Figure 3.5, Figure 3.6 shows the relative increase, R say, in regret for five different methods, with R defined as $R = (a-b)/b$, where a equals the regret when the empirically chosen version of h_1 is used, and b is the regret for the optimal choice of h_1. This comparison shows that the rule based on $\widehat{\text{err}}_{\mathcal{A}_1}$ performs better than the other four approaches; that cross-validation and the .632+ bootstrap perform worst; and that bagged cross-validation and Efron's (1983) approach are between those two groups.

We carried out the same experiment for the case where the normal mixture is replaced by a mixture of two lognormal distributions, or a mixture of two Cauchy distributions. In each setting the components in the mixture were taken to have the same location and scale parameters as in the normal case. The results are presented in Figure 3.7, and closely reflect those in Figure 3.6. In particular, apart from the Cauchy mixture case with $n = 50$, cross-validation and .632+ bootstrap methods again give the highest regret ratios. In the Cauchy case, although generally not for lognormal data, Efron's bootstrap methods edges out the bootstrap approach suggested in section 2.

We also explored properties of cross-validation in problems where at least one of the densities f and g becomes increasingly complex as sample size increases. This setting favours cross-validation. For example, taking f to be the uniform density on $[0, 1]$ and $g(x) = 1 + \cos(2k\pi x)$, the value of R, in the case of cross-validation, decreases as k increases across a broad range. The ratio also decreases if k and n increase together, in particular if $k = \log n$.

This reflects the fact that cross-validation is essentially a global procedure; it performs well at estimating tuning parameters determined by global issues, but does poorly at estimating those parameters when they have only a local influence. If two densities cross at only a small number of points, then, since the performance

of a classifier is determined by properties of the densities close to those points, optimising the classifier is a distinctly local problem. Therefore, cross-validation performs poorly. However, as the number of crossing points increases, the problem becomes more global in nature, and cross-validation becomes more competitive.

4. TECHNICAL ARGUMENTS

4.1. Preliminary arguments for Theorems 2.1 and 2.2. We shall assume throughout that $r = 1$; the case of general r differs only in notational complexity. We shall write y_1 as simply y.

Let $\mathcal{I}$ denote a compact interval containing a unique point, y say, for which $\Delta(y) = 0$; then, in view of (2.4), $\Delta'(y) \neq 0$. Assume, without loss of generality, that $\Delta'(y) > 0$. Put $\lambda = (\log n)^2$. Since $\lambda/(\log n)^{1/2} \to \infty$ then it may be proved using Bernstein's inequality, and the Hölder continuity of K, that for each $C > 0$,

$$P\Big\{\widehat{\Delta}_{f,-i}(x) < 0 \text{ for all } x \in \mathcal{I} \text{ with } x \le y - h^2\lambda, \\ \text{and } \widehat{\Delta}_{f,-i}(x) > 0 \text{ for all } x \in \mathcal{I} \text{ with } x \ge y + h^2\lambda\Big\} = 1 - O\big(n^{-C}\big)\,, \quad (4.1)$$

and that the same result holds if we replace $\widehat{\Delta}_{f,-i}$ by $\widehat{\Delta}$. The left-hand side of (4.1) does not depend on choice of i, and so that results holds uniformly in $1 \le i \le m$.

Split a sum over $1 \le i \le m$ into two parts, the first over i such that $|X_i - y| \le h^2\lambda$ and the second over i such that $|X_i - y| > h^2\lambda$, denoted by $\sum_i'$ and $\sum_i''$, respectively. Write $\int'$ and $\int''$ for integration over values x in the regions defined by $|x - y| \le h^2\lambda$ and $|x - y| > h^2\lambda$, respectively. Let M_- be the number of indices i for which $X_i - y < -h^2\lambda$ and $X_i \in \mathcal{I}$, and write $\mathcal{R}_-$ to denote the set of points x for which $x - y < -h^2\lambda$ and $x \in \mathcal{I}$. Then, by (4.1) and its analogue for $\widehat{\Delta}$,

$$P\bigg[\frac{1}{m}\sum_i{}'' I\big\{\widehat{\Delta}_{f,-i}(X_i) < 0\,, X_i \in \mathcal{I}\big\} = \frac{M_-}{m}\bigg] = 1 - O\big(n^{-C}\big)\,, \quad (4.2)$$

$$P\bigg[\int{}'' I\big\{\widehat{\Delta}(x) < 0\,, x \in \mathcal{I}\big\}\, f(x)\,dx = \int_{\mathcal{R}_-} f(x)\,dx\bigg] = 1 - O\big(n^{-C}\big)\,, \quad (4.3)$$

for each $C > 0$.

Write $\mathcal{R}$ for the set of values x satisfying $|x - y| \le h^2\lambda$, let M denote the number of X_j's that are in $\mathcal{R}$, and conditional on M, let $X_1', \ldots, X_M'$ be independent and identically distributed random variables, independent too of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, with the distribution of X conditional on $X \in \mathcal{R}$. Put

$$\bar{f}(x) = \frac{1}{mh_1}\bigg\{\sum_j{}'' K\bigg(\frac{x - X_j}{h_1}\bigg) + \sum_{j=1}^{M} K\bigg(\frac{x - X_j'}{h_1}\bigg)\bigg\}\,,$$

$$S(x) = \frac{1}{mh_1} \left\{ \sum_j{}' K\Big(\frac{x - X_j}{h_1}\Big) - \sum_{j=1}^{M} K\Big(\frac{x - X'_j}{h_1}\Big) \right\}.$$

Then, $\hat{f} = \bar{f} + S$, and moreover, $\bar{f}$ has the same distribution as $\hat{f}$. At some points in our argument we shall replace $\hat{f}$ by $\bar{f}$, with S as a remainder, since doing so reduces dependence on those data X_j and Y_j that lie between $y - h^2\lambda$ and $y + h^2\lambda$, and therefore makes subsequent analysis simpler.

Note that $E\{S(x) \,|\, M\} = 0$, and that, conditional on M, $mh_1\, S(x)$ can be written as a sum of $2M$ independent random variables, each of which is bounded by $C_1\,(h\lambda)^2$ uniformly in $x \in \mathcal{R}$ and $h_1 \in \mathcal{H}$. (Here and below, $C_1, C_2, \ldots$ will denote positive constants. To derive the bound we have used the fact that K has two bounded derivatives, and also the property $K'(0) = 0$.) Furthermore, for some $C_3 > 0$, M is bounded by $C_3 mh^2\lambda$ with probability $1 - O(n^{-C_2})$ for each $C_2 > 0$. These properties may be used to prove that for each integer $k \geq 1$, and with probability $1 - O(n^{-C})$ for each $C > 0$,

$$\sup_{x \in \mathcal{R}} E\big\{|S(x)|^{2k} \,\big|\, M\big\} \leq C_4(k) \left\{nh^2\lambda\,(h\lambda)^4/(nh)^2\right\}^k = C_4(k)\left(h^9\lambda^5\right)^k.$$

It now follows from Markov's inequality, and the Hölder continuity of K, that for each $C, \epsilon > 0$,

$$P\left\{ \sup_{x \in \mathcal{R},\, h_1 \in \mathcal{H}} |S(x)| > h^{(9/2)-\epsilon} \right\} = O\big(n^{-C}\big)\,. \tag{4.4}$$

More simply, for C_1 sufficiently large and for all $C > 0$,

$$P\left[\sup_{x \in \mathcal{I},\, h_1 \in \mathcal{H}} \left\{|\hat{f}(x)| + |\bar{f}(x)|\right\} > C_1 \right] = O\big(n^{-C}\big)\,. \tag{4.5}$$

Combining (4.4) and (4.5), and using the fact that $\hat{f} = \bar{f} + S$, it can be shown that

$$\begin{aligned} \hat{f}_{-i}(x) &= \frac{1}{m_1 h_1} \sum_{j=1}^{m}{}^{(i)}\, K\Big(\frac{x - X_j}{h_1}\Big) = \frac{m}{m_1}\,\hat{f}(x) - \frac{1}{m_1 h_1}\, K\Big(\frac{x - X_i}{h_1}\Big) \\ &= \bar{f}(x) - \frac{1}{mh_1}\, K\Big(\frac{x - X_i}{h_1}\Big) + \Theta_{1i}(x)\,, \end{aligned} \tag{4.6}$$

where, for each $C, \epsilon > 0$, and with $k = 1$,

$$P\left\{ \max_{1 \leq i \leq m} \sup_{x \in \mathcal{R},\, h_1 \in \mathcal{H}} |\Theta_{ki}(x)| > h^{(9/2)-\epsilon} \right\} = O\big(n^{-C}\big)\,. \tag{4.7}$$

The analogue of (4.6) when we do not omit the ith data value, and with a similar but simpler proof, is:

$$\hat{f}(x) = \bar{f}(x) + \Theta_1(x)\,, \tag{4.8}$$

where in place of (4.6), and for $k = 1$,

$$P\Big\{ \sup_{x \in \mathcal{R},\, h_1 \in \mathcal{H}} |\Theta_k(x)| > h^{(9/2)-\epsilon} \Big\} = O\big(n^{-C}\big)\,. \tag{4.9}$$

Combining (4.6) and (4.8) with the version of (4.8) that applies to $\hat{g}$, rather than to $\hat{f}$; and defining $\bar{g}$ analogously to $\bar{f}$; we see that if we define $\bar{\Delta} = p\,\bar{f} - (1-p)\,\bar{g}$, we obtain:

$$\widehat{\Delta}_{f,-i}(X_i) = \bar{\Delta}(X_i) - (mh_1)^{-1}\,K(0) + \Theta_{2i}(X_i)\,, \tag{4.10}$$

$$\widehat{\Delta}(x) = \bar{\Delta}(x) + \Theta_2(x)\,, \tag{4.11}$$

where Θ_{2i} and Θ_2 satisfy (4.7) and (4.9), respectively. (Here and below it will be assumed that the suprema in (4.7) and (4.9) are taken over $h_1, h_2 \in \mathcal{H}$, rather than simply over $h_1 \in \mathcal{H}$.)

It can be deduced from the pairs of results ((4.2), (4.10)) and ((4.3), (4.11)) that:

$$\begin{aligned} &\frac{1}{m} \sum_{i=1}^{m} I\big\{\widehat{\Delta}_{f,-i}(X_i) < 0\,, X_i \in \mathcal{I}\big\} \\ &\qquad = \frac{M_-}{m} + D_1 + \frac{1}{m} \sum_{j}{}' I\Big\{\bar{\Delta}(X_j) - (mh_1)^{-1}\,K(0) + \Theta_{2j}(X_j) < 0\Big\}, \end{aligned} \tag{4.12}$$

$$\begin{aligned} &\int I\big\{\widehat{\Delta}(x) < 0\,, x \in \mathcal{I}\big\}\, f(x)\,dx \\ &\qquad = \int_{\mathcal{R}_-} f(x)\,dx + D_2 + \int{}' I\big\{\bar{\Delta}(x) + \Theta_2(x) < 0\big\}\, f(x)\,dx\,, \end{aligned} \tag{4.13}$$

where $0 \le D_k \le 1$ and, for each $C > 0$,

$$P(D_k = 0) = 1 - O\big(n^{-C}\big)\,. \tag{4.14}$$

The next step is to show that the terms $\Theta_{2j}(X_j)$ and $\Theta_2(x)$, in (4.12) and (4.13), can in effect be replaced by deterministic quantities. If we replace $\Theta_{2j}(X_j)$ by $h^{(9/2)-\epsilon}$ or by $-h^{(9/2)-\epsilon}$ then, with probability equal to $1 - O(n^{-C})$ for all $C > 0$

(see (4.7)), we decrease the value of the series in (4.12), or increase it, respectively. The absolute value of the difference between these two versions of the series equals

$$S_1(h_1,h_2) \equiv \frac{1}{m}\sum_j{}' I\Big\{\Big|\bar\Delta(X_j) - (mh_1)^{-1}\,K(0)\Big| \le h^{(9/2)-\epsilon}\Big\}\,.$$

Define $\mu = E(\widehat\Delta) = E(\bar\Delta)$, $S_2(x,h_1,h_2) = \bar\Delta(x) - \mu(x)$ and $S_3 = S_2/h^2$. By Markov's inequality, using the fact that the number of summands in the series $\sum_j'$ is, with probability $1 - O(n^{-C})$ for each $C > 0$, bounded by $C_1 m h^2 \lambda$ for a constant $C_1 > 0$, and that f is bounded by a constant C_2 on the set $\mathcal{R}$, we deduce that, for all $C > 0$,

$$\begin{aligned}
e_n &\equiv E\Big\{\sup_{h_1,h_2\in\mathcal{H}} S_1(h_1,h_2)\Big\} \\
&\le E\left[\frac{1}{m}\sum_j{}' I\Big\{\inf_{h_1,h_2\in\mathcal{H}}\Big|\bar\Delta(X_j) - (mh_1)^{-1}\,K(0)\Big| \le h^{(9/2)-\epsilon}\Big\}\right] \\
&\le C_1\,C_2\,h^2\,\lambda\,\sup_{x\in\mathcal{R}} P\Big\{\inf_{h_1,h_2\in\mathcal{H}}\Big|\mu(x) + S_2(x,h_1,h_2) \\
&\qquad\qquad - (mh_1)^{-1}\,K(0)\Big| \le h^{(9/2)-\epsilon}\Big\} + O\big(n^{-C}\big)\,.
\end{aligned}$$

Now, $|\mu(x)| \le C_3\,h^2\lambda$, uniformly in $x \in \mathcal{R}$. Therefore,

$$e_n \le C_4\,h^2\,\lambda \sup_{|v|\le 2C_3\lambda,\,x\in\mathcal{R}} P\Big\{\inf_{h_1,h_2\in\mathcal{H}}\big|v + S_3(x,h_1,h_2)\big| \le h^{(5/2)-\epsilon}\Big\} + O\big(n^{-C}\big)\,.$$

The quantity $S_3(\,\cdot\,,h_1,h_2)$ has the same distribution as $h^{-2}\,(\widehat\Delta - \mu)$. Approximating to the latter using the approach of Komlós, Major and Tusnády (1976), and developing a concentration inequality for the resulting Gaussian process indexed by $h_1, h_2 \in \mathcal{H}$, we have:

$$P\Big\{\inf_{h_1,h_2\in\mathcal{H}}\big|v + S_3(x,h_1,h_2)\big| \le h^{(5/2)-\epsilon}\Big\} = O\big(h^{(5/2)-\epsilon-\delta}\big)\,,$$

uniformly in $|v| \le 2C_3\lambda$ and $x \in \mathcal{R}$, for each $\delta > 0$. Therefore, $e_n = O(h^{(9/2)-\epsilon-\delta}\,\lambda)$ for each $\delta > 0$. It follows that, for each $\epsilon > 0$,

$$\frac{1}{m}\sum_{i=1}^m I\big\{\widehat\Delta_{f,-i}(X_i) < 0\,, X_i \in \mathcal{I}\big\}$$

lies with probability 1 between the two values of

$$\frac{M_-}{m} + \frac{1}{m}\sum_j{}' I\Big\{\bar\Delta(X_j) - (mh_1)^{-1}\,K(0) < \pm h^{(9/2)-\epsilon}\Big\} \pm R_1(h_1,h_2)\,, \qquad (4.15)$$

where the plus and minus signs are taken respectively, and the nonnegative random process $R_1(h_1, h_2)$ satisfies, in the case $k = 1$,

$$E\Big\{ \sup_{h_1,h_2\in\mathcal{H}} |R_k(h_1, h_2)| \Big\} = O(h^{(9/2)-\epsilon-\delta}) \tag{4.16}$$

for each $\delta > 0$.

The same argument leads to the analogous result for the process of exceedences of zero by $\widehat{\Delta}_{g,-i}(Y_i)$. In this way it can be shown that for each $\epsilon > 0$, the cross-validation criterion $\mathrm{CV}(h_1, h_2)$ at (2.3) lies with probability 1, for all $h_1, h_2 \in \mathcal{H}$, between the two values of

$$\begin{aligned} &p\left[\frac{M_-}{m} + \frac{1}{m}\sum_j{}' I\Big\{\bar{\Delta}(X_j) - (mh_1)^{-1}\,K(0) < \pm h^{(9/2)-\epsilon}\Big\}\right] \\ &+ (1-p)\left[\frac{N_+}{n} + \frac{1}{n}\sum_j{}' I\Big\{\bar{\Delta}(Y_j) - (nh_2)^{-1}\,K(0) > \mp h^{(9/2)-\epsilon}\Big\}\right] \pm R_2(h_1, h_2)\,, \end{aligned} \tag{4.17}$$

where the plus and minus signs are taken respectively, and the nonnegative random process $R_2(h_1, h_2)$ satisfies (4.16) in the case $k = 2$ and for each $\delta > 0$. In (4.17), N_+ equals the number of Y_j's for which $Y_j - y > h^2\lambda$, the two summations are over j for which $|X_j - y| \le h^2\lambda$ and $|Y_j - y| \le h^2\lambda$, respectively, M and N are the respective numbers of such indices, and, conditional on M and N, $\bar{\Delta}$ is independent of the random variables X_j and Y_j for the indices j over which the summations are taken.

Similarly, using (4.13) in place of (4.12), it can be proved that for each $\epsilon > 0$, $\mathrm{emperr}_{\mathcal{A}_1}$ lies with probability 1, for all $h_1, h_2 \in \mathcal{H}$, between the two values of

$$\begin{aligned} &p\left[\int_{\mathcal{R}_-} f(x)\,dx + \int' I\big\{\bar{\Delta}(x) < \pm h^{(9/2)-\epsilon}\big\}\, f(x)\,dx\right] \\ &+ (1-p)\left[\int_{\mathcal{R}_+} g(x)\,dx + \int' I\big\{\bar{\Delta}(x) > \mp h^{(9/2)-\epsilon}\big\}\, g(x)\,dx\right] \pm R_3(h_1, h_2)\,, \end{aligned} \tag{4.18}$$

where $\mathcal{R}_-$ and $\mathcal{R}_+$ denote the sets of x for which $x - y > h^2\lambda$ and $x - y < -h^2\lambda$, respectively, and the nonnegative random process $R_3(h_1, h_2)$ satisfies (4.16).

Define $u_j = h_j/h$ for $j = 1, 2$, put $u = (u_1, u_2)$ and let $d(y \,|\, u)$ be as at (2.6). Since f has four bounded derivatives in an open neighbourhood of y, then Taylor expansion gives:

$$\mu(x) = E\{\widehat{\Delta}(x)\} = \Delta(x) + h^2\, d(y \,|\, u) + O\big(h^4\big)\,,$$

uniformly in values x for which $x = y + h^2 v$ where $|v| \le \lambda$, and in $h_1, h_2 \in \mathcal{H}$. Moreover, $\Delta(x) = h^2\, v\, \Delta'(y) + O(h^4\lambda^2)$, uniformly in $x \in \mathcal{R}$. Hence, for some $\epsilon > 0$,

$$\mu(x) = h^2 \{v\, \Delta'(y) + d(y \,|\, u)\} + O\big(h^4\lambda^2\big)\,, \tag{4.19}$$

uniformly in $x \in \mathcal{R}$ and in $h_1, h_2 \in \mathcal{H}$.

Define $D = h^{-2}\,(\bar\Delta - \mu)$. If we repeat the arguments in the paragraphs containing (4.15) and (4.17), but seek only a remainder of $h^{4-\epsilon}$, rather than $h^{(9/2)-\epsilon}$ as at present, then, for sufficiently large n, we can absorb the terms $(mh_1)^{-1}\,K(0)$ and $(nh_2)^{-1}\,K(0)$ in (4.15) and (4.17) into the remainders $\pm h^{4-\epsilon}$ and $\mp h^{4-\epsilon}$. Likewise, in view of (4.19) we can replace $\bar\Delta(x)$ by $h^2\,\{D + v\,\Delta'(y) + d(y \,|\, u)\}$. Therefore, multiplying by h^{-2} on either side of the inequalities in the arguments of the indicator functions, we deduce that for each $\epsilon > 0$, $\mathrm{CV}(h_1, h_2)$ lies with probability 1, for all $h_1, h_2 \in \mathcal{H}$, between the two values of

$$\begin{aligned} &p\left[\frac{M_-}{m} + \frac{1}{m}\sum_j{}' I\Big\{D(X_j) + V_j\,\Delta'(y) + d(y \,|\, u) < \pm h^{2-\epsilon}\Big\}\right] \\ &+ (1-p)\left[\frac{N_+}{n} + \frac{1}{n}\sum_j{}' I\Big\{D(Y_j) + W_j\,\Delta'(y) + d(y \,|\, u) > \mp h^{2-\epsilon}\Big\}\right] \pm R_4(h_1, h_2)\,, \end{aligned} \tag{4.20}$$

where V_j and W_j are defined by $X_j = y + h^2 V_j$ and $Y_j = y + h^2 W_j$, and the nonnegative random process R_4 satisfies, in place of (4.16),

$$E\Big\{\sup_{h_1, h_2 \in \mathcal{H}} |R_k(h_1, h_2)|\Big\} = O(h^{4-\epsilon-\delta})\,. \tag{4.21}$$

4.2. Proof of Theorem 2.1. The identity (2.10) can be derived using the methods leading to Theorem 2.1 of Hall and Kang (2005). Result (2.11) then follows from that theorem, under the additional condition on f and g that they are bounded away from zero on an open interval containing $\mathcal{I}$. (This is not needed here, however, since the assumptions for Theorem 2.1 are more restrictive about choice of bandwidth.) Result (2.12) follows from standard properties of density estimators. Therefore we confine attention to deriving (2.9).

Recall that we assume $r = 1$ and write $y = y_1$. Define $\pi_X = P(|X - y| \le h^2\lambda)$ and

$$\delta_X(x) = \frac{M - m\pi_X}{m(1-\pi_X)}\left[\frac{1}{\pi_X}\,E\Big\{\frac{1}{h_1}\,K\Big(\frac{x - X}{h_1}\Big)\,I(|X - y| \le h^2\lambda)\Big\} - E\big\{\hat f(x)\big\}\right],$$

and take δ_Y to be the analogous quantity for Y. Uniformly in $x \in \mathcal{R}$, we have $E\{\hat{f}(x) - \hat{f}(y)\} = O(h^2\lambda)$ and

$$\left| E\left\{ K\left(\frac{x-X}{h_1}\right) I(|X-y| \le h^2\lambda) \right\} \right. \\ \left. - E\left\{ K\left(\frac{y-X}{h_1}\right) I(|X-y| \le h^2\lambda) \right\} \right| = O\big(h^2\,\lambda^2\,\pi_X\big)\,.$$

Furthermore, $\pi_X \to 0$ as $n \to \infty$, and $P(|M - m\pi_X|/m > h^{(7/2)-\epsilon}) = O(n^{-C})$ for each $\epsilon, C > 0$. Combining these results we conclude that, for each $C, \epsilon > 0$ and in the case $Z = X$,

$$P\Big\{ \sup_{x\in\mathcal{R}} |\delta_Z(x) - \delta_Z(y)| > h^{(9/2)-\epsilon} \Big\} = O\big(n^{-C}\big)\,. \tag{4.22}$$

A similar argument establishes the same result for $Z = Y$. Therefore, using (4.22), and defining $\nu(x) = E\{D(x) \,|\, M, N\}$ (a random function) and $Z_1 = \nu(y) = O_p\{(h\lambda)^{1/2}\}$, we have:

$$\nu(x) = h^{-2}\,\{p\,\delta_X(x) - (1-p)\,\delta_Y(x)\} = Z_1 + \Psi_1(x)\,, \tag{4.23}$$

where, for $k = 1$ and for each $\epsilon, C > 0$,

$$P\Big\{ \sup_{x\in\mathcal{R}} |\Psi_k(x)| > h^{2-\epsilon} \Big\} = O\big(n^{-C}\big)\,. \tag{4.24}$$

Put $\bar{D} = D - E(D \,|\, M, N) = (1 - E')\,D$, where E' denotes the operator that takes expected value conditional on M and N. Write K' for the derivative of K, and define

$$Z_2 = (1-E') \left[\frac{p}{mh_1^2} \left\{ \sum_j{}'' K'\left(\frac{y-X_j}{h_1}\right) + \sum_{j=1}^{M} K'\left(\frac{y-X_j'}{h_1}\right) \right\} \right. \\ \left. - \frac{1-p}{mh_2^2} \left\{ \sum_j{}'' K'\left(\frac{y-Y_j}{h_2}\right) + \sum_{j=1}^{N} K'\left(\frac{y-Y_j'}{h_2}\right) \right\} \right] = O_p(h)\,,$$

where X_j' and Y_j' are the new, independent data added to replace those among the original data X_j and Y_j, respectively, that lie in $\mathcal{R}$; see the paragraph below that which contains (4.3). Taylor-expanding $D(y+h^2v)$ about $D(y)$, with $x = y+h^2v$ and the quadratic term as remainder, it can be shown that, with Ψ_2 satisfying (4.24),

$$\bar{D}(x) = \bar{D}(y) + v\,Z_2 + \Psi_2(x)\,. \tag{4.25}$$

Combining (4.23) and (4.25) we deduce that

$$D(x) = \bar{D}(x) + \nu(x) = \bar{D}(y) + Z_1 + v\,Z_2 + \Psi_3(x)\,, \tag{4.26}$$

where $\Psi_3 = \Psi_1 + \Psi_2$ and satisfies (4.24). Using (4.26) to substitute for D in (4.20), we deduce that for each $\epsilon > 0$, $\mathrm{CV}(h_1, h_2)$ lies with probability 1, for all $h_1, h_2 \in \mathcal{H}$, between the two values of

$$p\left(\frac{M_-}{m} + \frac{1}{m}\sum_j{}' I\left[V_j < \frac{\pm h^{2-\epsilon} - \{\bar{D}(y) + Z_1 + d(y\,|\,u)\}}{\Delta'(y) + Z_2}\right]\right)$$
$$+ (1-p)\left(\frac{N_+}{n} + \frac{1}{n}\sum_j{}' I\left[W_j > \frac{\mp h^{2-\epsilon} - \{\bar{D}(y) + Z_1 + d(y\,|\,u)\}}{\Delta'(y) + Z_2}\right]\right) \pm R_5(h_1, h_2)\,. \tag{4.27}$$

where the nonnegative random process R_5 satisfies (4.21) for each $\delta > 0$. Here we have used the fact that, with probability equal to $1 - O(n^{-C})$ for each $C > 0$, $|Z_2| \le \frac{1}{2}\,\Delta'(y)$ for all $h_1, h_2 \in \mathcal{H}$.

Let $\mathcal{F}$ denote the sigma-field generated by M, N, those values of X_i and Y_i that lie outside $\mathcal{R}$, and the added data X_i' and Y_i'. (There are just M X_i''s and N Y_i''s.) Put $A(v) = (v + \lambda)/2\lambda$. Conditional on $\mathcal{F}$, the variables V_j that are summed over in the first series in (4.27) are independent and identically distributed with distribution function

$$F_1(v) = \frac{F(y + h^2 v) - F(y - h^2\lambda)}{F(y + h^2\lambda) - F(y - h^2\lambda)} = A(v) + O\big(h^2\lambda\big)$$

for $|v| \le \lambda$, where the remainder is of the stated size uniformly in v in this range.

Define

$$Q_1 = -\frac{\bar{D}(y) + Z_1 + d(y\,|\,u)}{\Delta'(y) + Z_2}\,, \quad Q_2 = -\frac{\bar{D}(y) + d(y\,|\,u)}{\Delta'(y)}\,,$$

and observe that for each $\delta > 0$,

$$\begin{aligned} Q_0 &\equiv \frac{\pm h^{2-\epsilon} - \{\bar{D}(y) + Z_1 + d(y\,|\,u)\}}{\Delta'(y) + Z_2} \\ &= Q_1 + O_p\big(h^{2-\epsilon}\big) = Q_2 + O_p\big(h^{2-\epsilon} + h^{(1/2)-\delta}\big)\,, \end{aligned} \tag{4.28}$$

uniformly in $h_1, h_2 \in \mathcal{H}$.

Let $\mathcal{W}$ be a standard Wiener process, and put $\mathcal{B}(t) = \mathcal{W}(t) - t\,\mathcal{W}(1)$, a Brownian bridge. Define

$$\mathcal{W}_{1X}(t) = (2\lambda)^{1/2}\left[\mathcal{W}\big\{\tfrac{1}{2} + (2\lambda)^{-1}\,t\big\} - \mathcal{W}\big(\tfrac{1}{2}\big)\right],$$

a second standard Wiener process. In this notation,

$$(2\lambda)^{1/2}\left[\mathcal{B}\{\tfrac{1}{2}+(2\lambda)^{-1}t\}-\mathcal{B}\big(\tfrac{1}{2}\big)\right]=\mathcal{W}_{1X}(t)-(2\lambda)^{-1/2}\,t\,\mathcal{W}(1)\,. \tag{4.29}$$

Condition on $\mathcal{F}$, and hence also on $\bar{D}(y)$ and therefore also on Q_2. Replacing t by Q_2 in (4.29), and using a stochastic approximation to sums of independent random variables, we see that we may choose $\mathcal{B}$, and hence $\mathcal{W}_{1X}$, such that, uniformly in $h_1, h_2 \in \mathcal{H}$,

$$\begin{aligned}
&\frac{1}{m}\sum_j{}'\{I(V_j<Q_0)-I(V_j<0)\}\\
&=\frac{M}{m}\,\{F_1(Q_0)-F_1(0)\}+\frac{M^{1/2}}{m}\,\Big[\mathcal{B}\{F_1(Q_0)\}-\mathcal{B}\{F_1(0)\}\Big]+o_p\big(h^{7/2}\big)\\
&=\frac{M}{m}\,\{F_1(Q_0)-F_1(0)\}+\frac{M^{1/2}}{m}\,\Big\{\mathcal{B}\Big(\frac{Q_0+\lambda}{2\lambda}\Big)-\mathcal{B}\big(\tfrac{1}{2}\big)\Big\}+o_p\big(h^{7/2}\big)\\
&=\frac{M}{m}\,\{F_1(Q_1)-F_1(0)\}+\frac{M^{1/2}}{m}\,\Big\{\mathcal{B}\Big(\frac{Q_2+\lambda}{2\lambda}\Big)-\mathcal{B}\big(\tfrac{1}{2}\big)\Big\}+o_p\big(h^{7/2}\big)\\
&=\frac{M}{m}\,\{F_1(Q_1)-F_1(0)\}+\big\{2\,m^{-1}\,h^2\,\lambda\,f(y)\big\}^{1/2}\,(2\lambda)^{-1/2}\,\mathcal{W}_{1X}(Q_2)+o_p\big(h^{7/2}\big)\\
&=\frac{M}{m}\,\frac{Q_1}{2\lambda}+\big\{m^{-1}\,h^2\,f(y)\big\}^{1/2}\,\mathcal{W}_{1X}(Q_2)+o_p\big(h^{7/2}\big)\,.
\end{aligned} \tag{4.30}$$

To obtain the third identity here we employed (4.28) and the fact that $M = O_p(mh^2\lambda)$; for the fourth identity, we noted that $M/\{2mh^2\lambda\,f(y)\}$ converges to 1 in probability, and used the fact that $\lambda = (\log n)^2$; and to get the final identity we used (4.29).

Similarly, we may construct, conditional on $\mathcal{F}$, a Wiener process $\mathcal{W}_{1Y}$ that is independent of $\mathcal{W}_{1X}$, such that, uniformly in $h_1, h_2 \in \mathcal{H}$,

$$\begin{aligned}
&\frac{1}{n}\sum_j{}'\left\{I\left[W_j>\frac{\mp h^{2-\epsilon}-\{\bar{D}(y)+Z_1+d(y\,|\,u)\}}{\Delta'(y)+Z_2}\right]-I(W_j>0)\right\}\\
&\qquad=-\frac{N}{n}\,\frac{Q_1}{2\lambda}+\big\{n^{-1}\,h^2\,g(y)\big\}^{1/2}\,\mathcal{W}_{1Y}(Q_2)+o_p\big(h^{7/2}\big)\,.
\end{aligned} \tag{4.31}$$

In both (4.30) and (4.31) we can replace Q_2 by $-Q_3$, where

$$Q_3=\frac{D_0(y)+d(y\,|\,u)}{\Delta'(y)}\,, \tag{4.32}$$

where $D_0 = -h^{-2}\,(1-E)\,\widehat{\Delta}$ and E is the expectation operator. Since $Q_2 = -Q_3 + O_p(h^\delta)$, uniformly in $h_1, h_2 \in \mathcal{H}$, for $\delta > 0$ sufficiently small, then the remainder after the alteration to (4.30) and (4.31) goes into the $o_p(h^{7/2})$ remainder terms.

With this modification, combining (4.30) and (4.31) we deduce that the quantity displayed at (4.27) can be written as $T_2+T_3+o_p(h^{7/2})$, uniformly in $h_1, h_2 \in \mathcal{H}$, where

$$T_1 = \frac{p\,M^-}{m} + \frac{(1-p)\,N^+}{n}\,, \quad T_2 = T_1 + \left\{\frac{p\,M}{m} - \frac{(1-p)\,N}{n}\right\}\frac{Q_1}{2\lambda}\,,$$
$$T_3 = h\left[p\,\{f(y)/m\}^{1/2}\,\mathcal{W}_X(Q_3) + (1-p)\,\{g(y)/n\}^{1/2}\,\mathcal{W}_Y(Q_3)\right], \qquad (4.33)$$

$\mathcal{W}_Z(t) = \mathcal{W}_{1Z}(-t)$ for $Z = X$ and $Z = Y$, and M^- (respectively, N^+) is the number of indices j such that $X_j < y$ ($Y_j > y$). Therefore,

$$\mathrm{CV}(h_1, h_2) = T_2(h_1, h_2) + T_3(h_1, h_2) + o_p\big(h^{7/2}\big)\,, \qquad (4.34)$$

uniformly in $h_1, h_2 \in \mathcal{H}$.

Note particularly that the Wiener processes $\mathcal{W}_X$ and $\mathcal{W}_Y$ are constructed to be independent conditional on $\mathcal{F}$. The variable Q_3 is measurable in $\mathcal{F}$, and so, in the definition of T_3 at (4.33), the processes $\mathcal{W}_X$ and $\mathcal{W}_Y$, and the random variable Q_3, are all independent.

Standard calculations show that, with $R = p\,(M/m) - (1-p)\,(N/n)$, we have:

$$E(R) = \int_{-h^2\lambda}^{h^2\lambda} \Delta(y+u)\,du = O\big\{(h^2\lambda)^2\big\}\,, \qquad \sup_{h_1,h_2\in\mathcal{H}} |Q_1(h_1,h_2)| = O_p\big\{(\log n)^{1/2}\big\}$$

and $\mathrm{var}(R) = O(h^7\lambda)$. Moreover, R depends on neither h_1 nor h_2. From these properties, recalling that $\lambda = (\log n)^2$, we deduce that $T_2 - T_1 = o_p(h^{7/2})$, uniformly in $h_1, h_2 \in \mathcal{H}$. Combining these results with (4.34) we deduce that

$$\mathrm{CV}(h_1, h_2) = T_1(h_1, h_2) + T_3(h_1, h_2) + o_p\big(h^{7/2}\big)\,, \qquad (4.35)$$

uniformly in $h_1, h_2 \in \mathcal{H}$.

Standard strong approximation arguments show that the stochastic process $Q_3(u_1, u_2)$ can be written as

$$Q_3(u_1,u_2)\,\Delta'(y) = p\,\{\rho\,f(y)\}^{1/2}\,\mathcal{G}_X(u_1) + (1-p)\,g(y)^{1/2}\,\mathcal{G}_Y(u_2) + d(y\,|u) + o_p(1)$$
$$= \mathcal{V}(u_1,u_2) + o_p(1)\,, \qquad (4.36)$$

where $\mathcal{G}_X$ and $\mathcal{G}_Y$ are independent and identically distributed Gaussian processes, each with zero mean and covariance given by (2.5), and $\mathcal{V}$ has the same definition as $\mathcal{V}_i$, at (2.7), but with y_i there replaced by y here. The desired result (2.9), in the case $r = 1$ and with $\Delta'(y_1) = \Delta'(y) > 0$, follows from (4.35) and (4.36).

4.3. Proof of Theorem 2.2. The method of proof here is similar to that given in section 4.2. There, starting from the result given in the paragraph containing (4.17), we derived first the result in the paragraph containing (4.27), and from there we went to (4.35). On the present occasion we replace the result in the paragraph containing (4.17) by the result in the paragraph containing (4.18), and in this way show that the quantity

$$\text{emperr}_{\mathcal{A}_1} - \text{err}_{\mathcal{A}_1} = \int \Big[I\big\{\widehat{\Delta}(x) < 0\big\} - P\big\{\widehat{\Delta}(x) < 0\big\} \Big]\, \Delta(x)\, dx$$

lies, with probability 1, between the two values of

$$\int{}' \Big[I\big\{\bar{\Delta}(x) < \pm h^{(9/2)-\epsilon}\big\} - P\big\{\bar{\Delta}(x) < \mp h^{(9/2)-\epsilon}\big\} \Big]\, \Delta(x)\, dx \pm R_6(h_1, h_2)\,, \quad (4.37)$$

where the nonnegative random process $R_6(h_1, h_2)$ satisfies (4.16).

Changing variable from $x = y + h^2 v$ to v, and using the argument leading from the result in the paragraph containing (4.27) to (4.35), we find that both of the two quantities in (4.37) equal

$$\begin{aligned} h^4\, \Delta'(y) \int_{|v| \le \lambda} \Big[I\{v < -Q_3/\Delta'(y)\} - P\{v < -Q_3/\Delta'(y)\} \Big]\, dv + o_p\big(h^4\big) \\ = h^4\, \Delta'(y) \int_{|v| \le \lambda} \Big[I\{v < \mathcal{V}(u)\} - P\{v < \mathcal{V}(u)\} \Big]\, dv + o_p\big(h^4\big)\,, \end{aligned}$$

uniformly in $h_1, h_2 \in \mathcal{H}$, where Q_3 is as at (4.32). This result leads directly to (2.13).

REFERENCES

ANCUKIEWICZ, M. (1998). An unsupervised and nonparametric classification procedure based on mixtures with known weights. *J. Classification* **15**, 129–141.

BOWMAN, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360.

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

CHANDA, K.C. AND RUYMGAART, F.H. (1989). Asymptotic estimate of probability of misclassification for discriminant rules based on density estimates. *Statist. Probab. Lett.* **8**, 81–88.

DEVROYE, L., GYÖRFI, L. AND LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316–331.

EFRON, B. AND TIBSHIRANI, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92**, 548–560.

FARAWAY, J.J. AND JHUN, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85**, 1119–1122.

FIX, E. AND HODGES, J. (1951). Discriminatory analysis. Nonparametric discrimination: consistency properties. Technical Report No. 4, Project No. 21–49–004, USAF School of Aviation Medicine, Randolph Field, TX.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11**, 1156–1174.

HALL, P. AND KANG, K.-H. (2005). Bandwidth choice for nonparametric classification. *Ann. Statist.* **33**, 284–306.

KOMLÓS, J., MAJOR, P. AND TUSNÁDY, G. (1976). An approximation of partial sums of independent rv's, and the sample df. II. *Z. Wahrscheinlichkeitstheorie verv. Geb.*, **34**, 33–58.

KRZYŻAK, A. (1991). On exponential bounds on the Bayes risk of the nonparametric classification rules. In: *Nonparametric Functional Estimation and Related Topics* (Spetses, 1990), *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335**, 347–360.

LAPKO, A.V. (1993). *Nonparametric Classification Methods and their Application.* (In Russian.) VO Nauka, Novosibirsk.

LUGOSI, G. AND NOBEL, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24**, 687–706.

LIN, C.T. (2001). Nonparametric classification on two univariate distributions. *Commun. Statist. Theory Meth.* **30**, 319–330.

LUGOSI, G. AND PAWLAK, M. (1994). On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Trans. Inform. Theory* **40**, 475–481.

MAMMEN, E. AND TSYBAKOV, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808–1829.

PAWLAK, M. (1993). Kernel classification rules from missing data. *IEEE Trans. Inform. Theory* **39**, 979–988.

STEELE, B.M. AND PATTERSON, D.A. (2000). Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: Applications for classification and error assessment. *Statistics and Computing* **10**, 349–355.

STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 1285–1297.

YANG, Y.H. (1999a). Minimax nonparametric classification — Part I: Rates of convergence. *IEEE Trans. Inform. Theory* **45**, 2271–2284.

YANG, Y.H. (1999b). Minimax nonparametric classification — Part II: Model selection for adaptation. *IEEE Trans. Inform. Theory* **45**, 2285-2292.

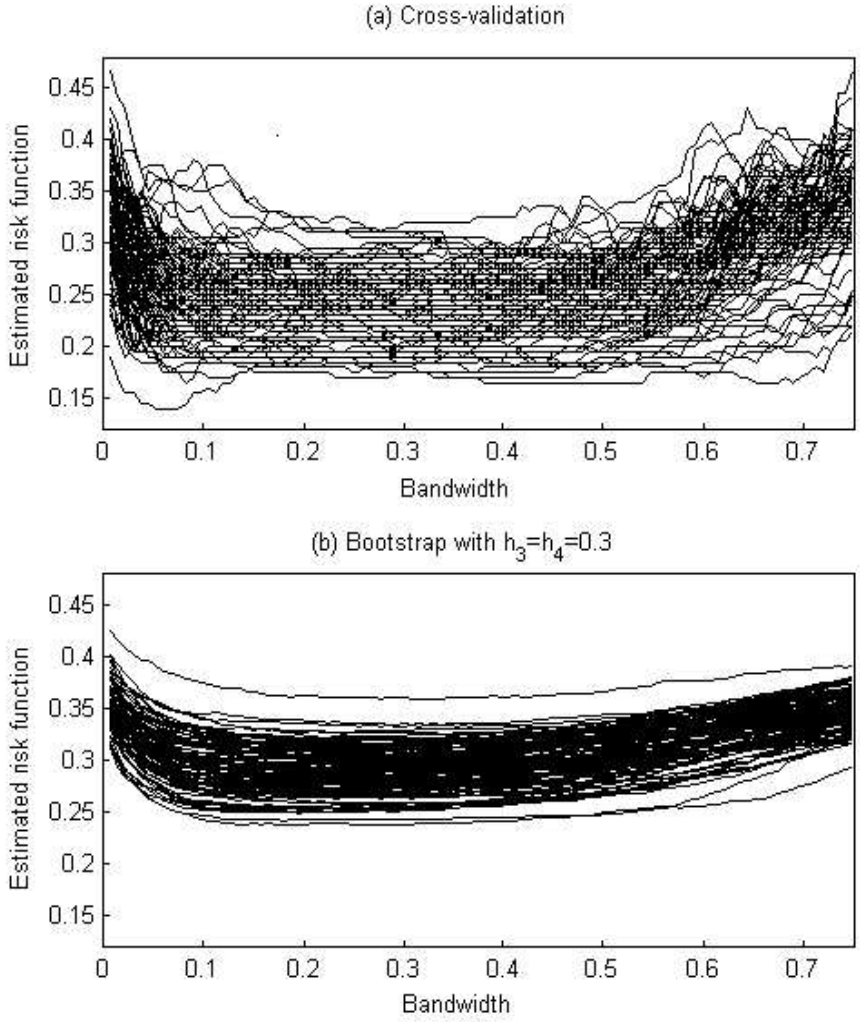

Figure 3.1 : Estimated risk using (a) cross-validation method or (b) the bootstrap.

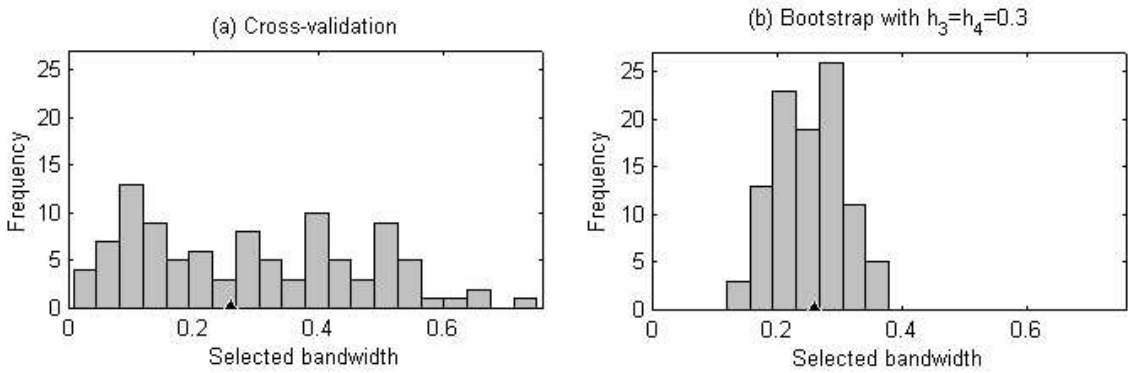

Figure 3.2 : Histogram estimators of empirical bandwidth distributions, when bandwidths are selected using (a) cross-validation or (b) the bootstrap.

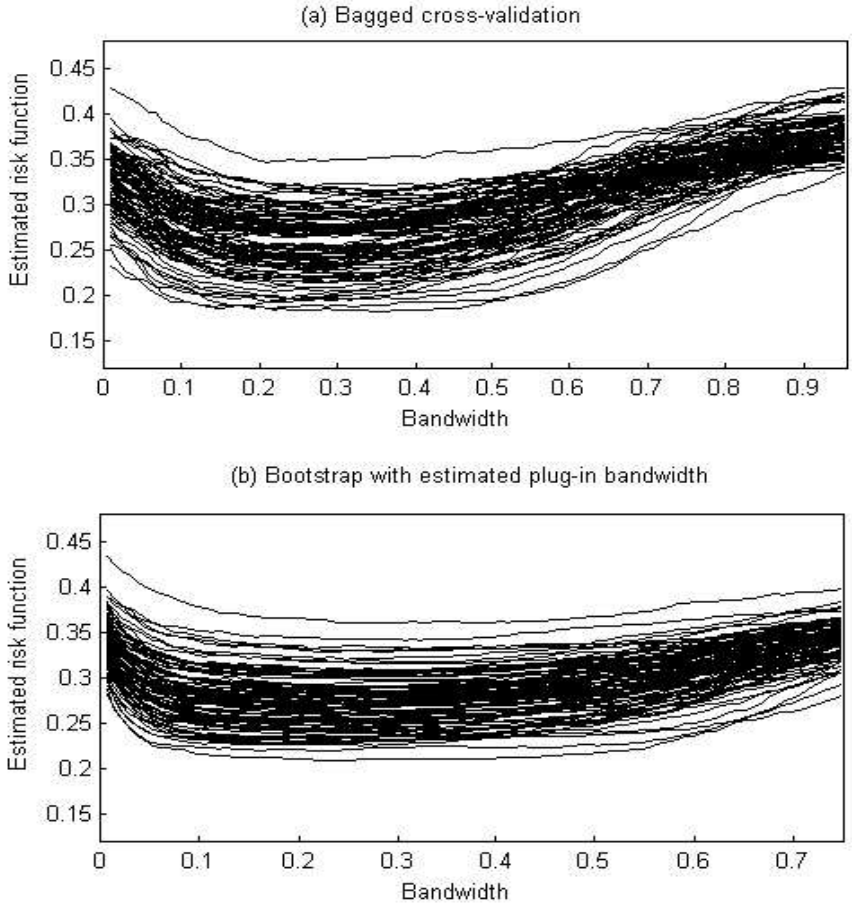

Figure 3.3 : Estimated risk for (a) bagged version of cross-validation or (b) bootstrap using estimated plug-in bandwidths.

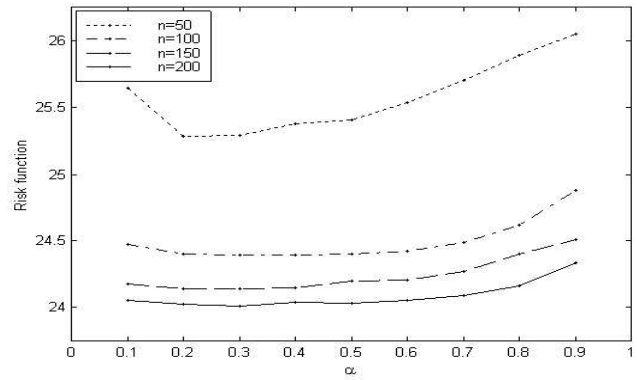

Figure 3.4 : Average risk for the bagged cross-validation rule, expressed as a function of α, equal to the proportion of the sample that is resampled.

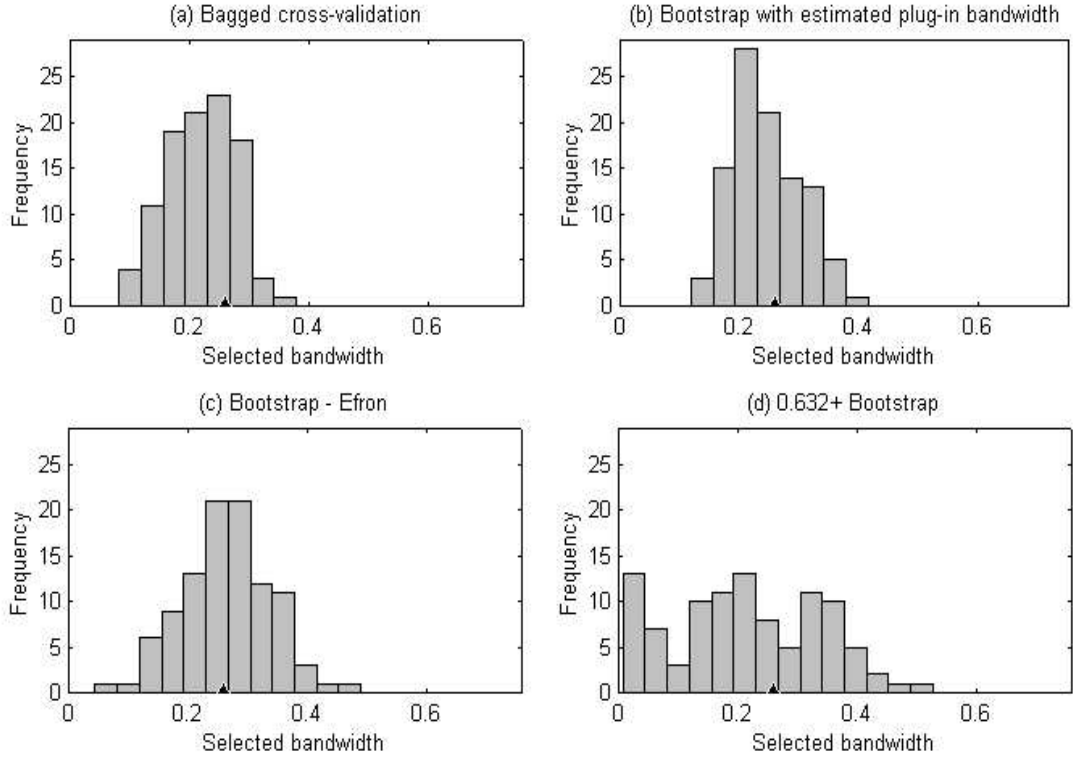

Figure 3.5 : Histogram estimators of empirical bandwidth distributions, when bandwidths are selected using (a) bagged cross-validation, (b) bootstrap employing estimated plug-in bandwidths, (c) Efron's (1983) method, or (d) the .632+ bootstrap.

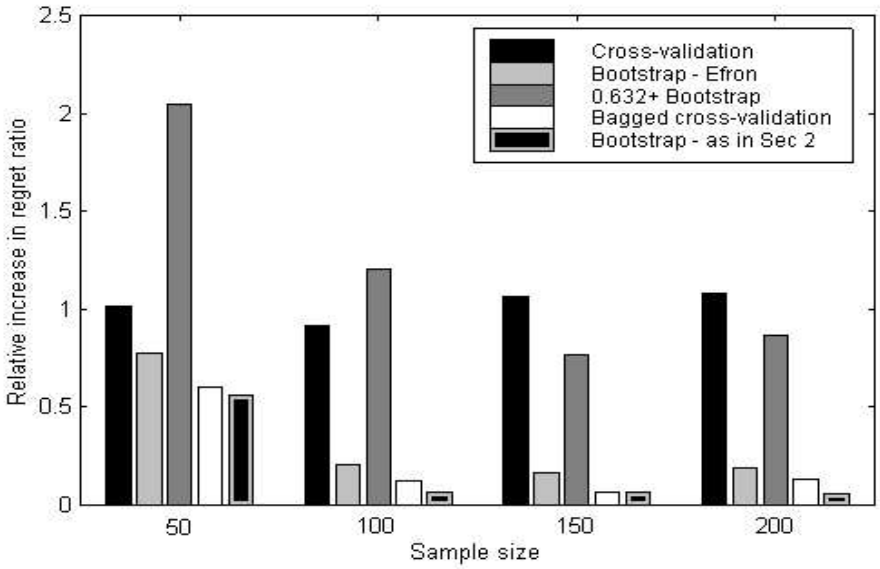

Figure 3.6 : Relative increase in regret for different classifiers, in the case of normal mixture distributions.

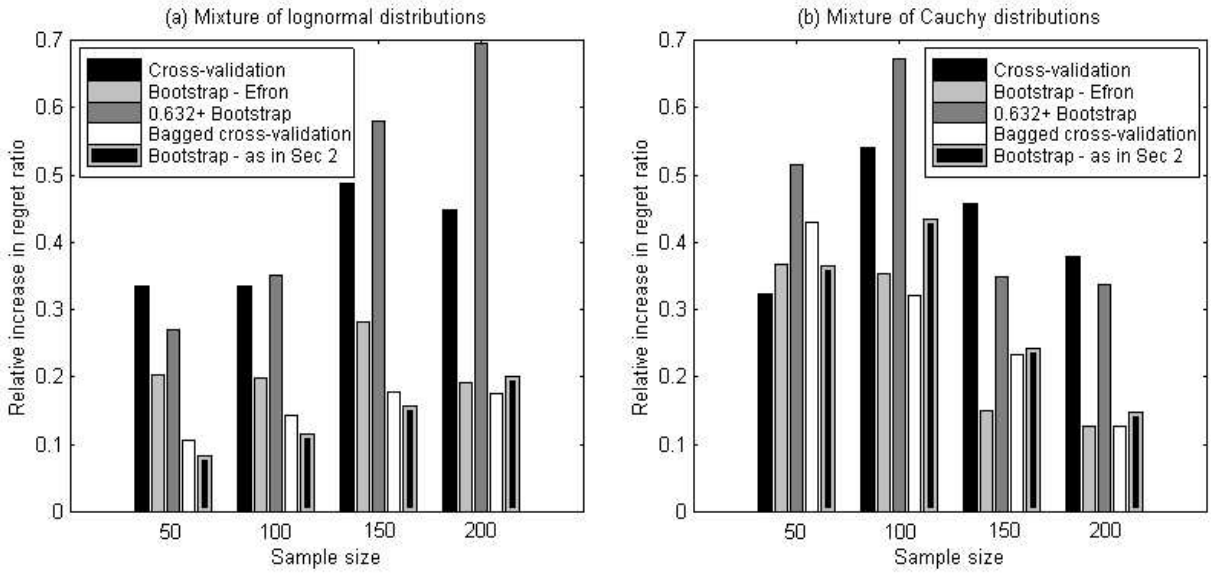

Figure 3.7 : Relative increase in regret for different classifiers, in the case of (a) log-normal mixtures or (b) Cauchy mixtures.