CROSS-VALIDATION AND THE ESTIMATION OF CONDITIONAL PROBABILITY DENSITIES

Peter Hall[1] Jeff Racine[2] Qi Li[3]

ABSTRACT. Many practical problems, especially some connected with forecasting, require nonparametric estimation of conditional densities from mixed data. For example, given an explanatory data vector X for a prospective customer, with components that could include the customer's salary, occupation, age, sex, marital status and address, a company might wish to estimate the density of the expenditure, Y, that could be made by that person, basing the inference on observations of (X, Y) for previous clients. Choosing appropriate smoothing parameters for this problem can be tricky, not least because plug-in rules take a particularly complex form in the case of mixed data. An obvious difficulty is that there exists no general formula for the optimal smoothing parameters. More insidiously, and more seriously, it can be difficult to determine which components of X are relevant to the problem of conditional inference. For example, if the jth component of X is independent of Y then that component is irrelevant to estimating the density of Y given X, and ideally should be dropped before conducting inference. In this paper we show that cross-validation overcomes these difficulties. It automatically determines which components are relevant and which are not, through assigning large smoothing parameters to the latter and consequently shrinking them towards the uniform distribution on the respective marginals. This effectively removes irrelevant components from contention, by suppressing their contribution to estimator variance; they already have very small bias, a consequence of their independence of Y. Cross-validation also gives us important information about which components are relevant: the relevant components are precisely those which cross-validation has chosen to smooth in a traditional way, by assigning them smoothing parameters of conventional size. Indeed, cross-validation produces asymptotically optimal smoothing for relevant components, while eliminating irrelevant components by oversmoothing. In the problem of nonparametric estimation of a conditional density, cross-validation comes into its own as a method with no obvious peers.

KEYWORDS. Bandwidth choice, binary data, categorical data, continuous data, dimension reduction, discrete data, kernel methods, mixed data, nonparametric density estimation, relevant and irrelevant data, smoothing parameter choice.

SHORT TITLE. Cross-validation.

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

[2]Department of Economics and Center for Policy Research, Syracuse University, Syracuse, NY 13244-1020, USA

[3]Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA

1. INTRODUCTION

Conditional probability density functions play a key role in applied statistical analysis, particularly in economics. Such densities are especially important in prediction problems, where, for a given value of a vector X of explanatory variables, we wish to estimate the conditional density of a response, Y. From some viewpoints this is a conventional problem; both parametric and nonparametric methods for estimating conditional distributions already exist. However, the problem has the distinctive feature that if components of the vector X contain no information about Y, and are "irrelevant" in this sense to the problem of estimating the conditional density, then they should be dropped when conducting inference. Not doing so can seriously inhibit performance, since then conditional inference will be based on data whose dimension is too high, degrading both the mathematical convergence rate and the method's statistical accuracy.

When the conditional density is estimated nonparametrically, the problem of choosing "relevant" components among the explanatory variables is closely related to that of selecting smoothing parameters. For example, if X is p-variate, and has a continuous distribution, then a conventional estimator of the density $g(y|x)$ of Y given $X = x$, using second-order kernels and a sample of size n, converges at rate $n^{-2/(p+5)}$ (assuming Y is continuous). This rate is achieved using bandwidths of size $n^{-1/(p+5)}$. If, however, p_2 of those components are irrelevant to the problem of estimating the distribution of Y given X, for example because they are stochastically independent of Y, then we can remove them and improve the convergence rate to $n^{-2/(p_1+5)}$, where $p_1 = p - p_2$. To achieve this outcome, the size of bandwidth should be reduced to $n^{-1/(p_1+5)}$.

The result of reducing the length of X in this way is distinctly different from that achieved by more conventional dimension reduction methods, for example projection pursuit. The latter generally develops only an *approximation* to g; the approximation would generally not consistently estimate the true conditional density as n increased. By way of contrast, if we could identify components of X that were independent of Y then these would be deleted at the outset, leading to improvements in the accuracy with which the density of Y, given X, was consistently estimated.

In applications of smoothing methods to real data, in the context of estimating

conditional densities, we have found that "irrelevant" components are surprisingly common. See in particular the examples in section 5, based on two classic benchmark datasets. In principle this problem can be tackled by applying a battery of hypothesis tests, prior to conducting inference. Tests for independence of individual components, or of groups or linear combinations of components, can be used to identify irrelevant explanatory variables. However, such an approach is awkward and tedious to implement, not least because the components of X will often be of many different types — continuous, unordered discrete, and ordered discrete, all in the same vector. We shall suggest instead a version of cross-validation in this context, and show that it has virtues that make it especially suited to simultaneously choosing smoothing parameters and removing irrelevant components of explanatory variables.

To describe how cross-validation works in this problem, let us assume initially that all components of X are continuous; this will simplify exposition. Construction of the cross-validation criterion, CV say, is not trivial, but a certain weighted form of it has an elementary form. If p_2 of the components of X are independent of Y and therefore irrelevant, and if the remaining p_1 components of X are relevant, then the empirical bandwidths that minimise CV will demonstrate markedly dichotomous behaviour: those that correspond to irrelevant components will diverge to infinity as the sample size increases, whereas those that correspond to relevant components will consistently estimate the bandwidths, of size $n^{-1/(p_1+5)}$, that would be appropriate if only the relevant components were present.

By diverging to infinity, the bandwidths for irrelevant components effectively shrink those components to a distribution that is virtually "uniform on the real line," and so eliminate the irrelevant components from contention. Therefore, without any particular inputs being made from the experimenter, cross-validation automatically identifies relevant and irrelevant components, removes the latter, and chooses the correct bandwidths for the former. It conducts a *de facto* dimension reduction program, tailored to the problem of estimating the conditional density.

Similar behaviour is observed when one or more of the components of X are discrete: application of cross-validation to selecting smoothing parameters effectively shrinks each discrete irrelevant component to the uniform distribution on its support, thereby effectively removing it from contention in the problem of estimat-

ing the conditional density of Y given X. For the relevant components that remain, cross-validation automatically chooses smoothing parameters that are appropriate when only the relevant components are used for inference.

These results continue to hold in the case of conditional density estimation from explanatory data whose components are both continuous and discrete. For simplicity, in our theoretical work we shall treat only mixed unordered discrete and continuous components. However, when these are combined with ordered discrete components the results are virtually identical. Each context is characterised, in the case of irrelevant components, by divergence of smoothing parameters to the upper extremities of their respective ranges, or equivalently, by shrinkage to the uniform distribution.

In view of our focus on mixed data we shall address only the setting where a different smoothing parameter is used for each component. Similar results are obtained, however, if smoothing is done more comprehensively, for example by using a $p \times p$ bandwidth matrix to smooth p-variate explanatory variables where all components are continuous. In this case, cross-validation automatically identifies linear transformations of X that are independent of Y, and eliminates them by shrinking them to the uniform distribution on the real line.

The divergence of cross-validation smoothing parameters to their upper extremities, which characterises the case of irrelevant components, provides invaluable empirical advice about which components are relevant and which are not. It comes "for free" when we use cross-validation to select the amount of smoothing. A formal statistical test of independence would have greater power, but would be substantially more awkward to implement.

An alternative approach to solving this problem would be to use classical variable selection methods to choose relevant components, and employ a separate smoothing-parameter choice technique to determine how much smoothing to do. However, the fact that this problem involves both continuous and discrete variables means that there does not exist an off-the-shelf algorithm for choosing smoothing parameters, except for the cross-validation approach that we suggest. In particular, configuring plug-in rules for mixed data is an algebraically tedious task, and in fact no general formulae are available. Additionally, plug-in rules, even after adaptation to mixed data, require choice of "pilot" smoothing parameters, and it is not clear

how to best make that selection for the continuous and discrete variables involved. As we shall show, cross-validation avoids these problems and has the additional virtue of separating variables into relevant and irrelevant categories.

Our method can readily be generalised to cover other econometric models with mixed discrete and continuous variables. For example, Hahn (1998) and Hirano, Imbens and Ridder (2002) considered the nonparametric estimation of average treatment effects, Horowitz (2001) dealt with nonparametric estimation of a generalised additive model having an unknown link function, and Lewbel and Linton (2002) treated nonparametric censored and truncated regression models. Each of these approaches assumes that the nonparametric covariates are continuous (or, when discrete covariates are present, uses the nonparametric frequency method). One can employ the cross validation-based smoothing method presented in this paper to generalise the aforementioned approaches to handle mixed discrete and continuous covariates. Such an extension also has the advantage of being able to remove irrelevant covariates (both discrete and continuous), thereby yielding more reliable estimation results.

There is an alternative approach to defining relevance and irrelevance, based on conditional independence rather than conventional independence. To describe it, let us again assume for simplicity that all the components of X are continuous. We might say that $X = (X^{[1]}, X^{[2]})$ represents a decomposition of X into relevant and irrelevant parts $X^{[1]}$ and $X^{[2]}$, respectively, if Y and $X^{[2]}$ are independent conditional on $X^{[1]}$. While this approach is attractive in at least a theoretical sense, it has certain difficulties from an operational viewpoint. To appreciate why, consider (for example) the case where $Y = Z_1 + Z_2 + Z_3$, $X_j = Z_j + \epsilon Z_{j+3}$ for $j = 1, 2$, $X = (X_1, X_2)$, $\epsilon > 0$, and $Z_1, \ldots, Z_5$ are independent standard normal random variables. If ϵ is small then, depending on which of X_1 and X_2 we decide to condition on, a practical assessment of "relevance" that is based on conditional independence is likely to suggest that either X_1 or X_2, but not both, is irrelevant.

Therefore, in practical terms, and using an assessment based on conditional independence, the problem can be ambiguous, and empirical difficulties may be expected to arise when deciding how to partition X into relevant and irrelevant parts. On the other hand, if an unconditional view of independence is taken, as suggested in the present paper, then our method will generally conclude that both X_1 and X_2

are relevant, even for small ϵ. However, in cases where the sort of ambiguity mentioned above does not arise, sketched theoretical analyses in particular cases, and small scale simulation studies, suggest that cross-validation will successfully detect irrelevance, by virtue of the corresponding bandwidths diverging, when irrelevance is defined in the sense of conditional independence.

Section 2 introduces our cross-validation algorithm, and section 3 develops properties of mean squared error, and of optimal smoothing parameters, when no irrelevant components are present. The results there set theoretical benchmarks for performance of bandwidth selectors after irrelevant components have been removed. We show in section 4 that cross-validation attains these benchmarks. In that section we give concise, mathematical definitions of what we mean by "relevant" and "irrelevant" components. Numerical illustrations of the performance of cross-validation in removing irrelevant components, and conducting adaptive inference, are given in section 5. There we pay particular attention to the case of mixed continuous and discrete explanatory variables, and we also apply our method to two well known datasets having a large number of discrete cells relative to their sample sizes (therefore the conventional frequency method is infeasible for both datasets). We show that our proposed estimator smoothes out some irrelevant variables and yields better out-of-sample predictions than some commonly used parametric methods.

The use of least-squares cross-validation to select smoothing parameters in density estimation dates from work of Rudemo (1982) and Bowman (1984), following earlier discussion of the Kullback-Leibler case by Habbema, Hermans and Van Den Broek (1974). Tutz (1991) treated cross-validation for conditional density estimation from mixed variables. Theory for least-squares cross-validation was developed by Hall (1983b, 1985) and Stone (1984), and second-order properties were addressed by Hall and Marron (1987).

Smoothing methods for ordered categorical data have been surveyed by Simonoff (1996, section 6). Hall (1983a) and Li and Racine (2003) treated unconditional joint density estimation from mixed data. There is a large literature on dimension reduction for density estimation, including work of Friedman, Stuetzle and Schroeder (1984) and Jones and Sibson (1987).

One of the reasons for estimating conditional densities, rather than conditional distributions, is that they give a better idea of the relative placement of "weight"

in the distribution. As a result, there is constant, continuing interest in the topic of conditional density estimation. For a recent reference, see the work of Fan and Yim (2003), who discuss novel methods for conditional density estimation.

2. METHODOLOGY FOR CROSS-VALIDATION

Let $\hat{f}$ denote an estimator of the density, f, of (X, Y), and let $\widehat{m}$ be an estimator of the marginal density, m, of X. We estimate $g(y|x) = f(x, y)/m(x)$, the density of Y conditional on X, by $\hat{g}(y|x) = \hat{f}(x, y)/\widehat{m}(x)$, and use as our performance criterion the weighted integrated squared error,

$$\text{ISE} = \int \{\hat{g}(y|x) - g(y|x)\}^2 \, m(x) \, dW(x) \, dy \,, \tag{2.1}$$

where $dW(x)$ denotes the infinitesimal element of a measure.

The presence of $dW(x)$ at (2.1) serves only to avoid difficulties caused by dividing by zero, or by numbers close to zero, in the ratio $\hat{f}(x, y)/\widehat{m}(x)$. This is usually only a problem for the continuous components of x. Therefore, if X denotes a generic X_i, if $X = (X^{\mathrm{c}}, X^{\mathrm{d}})$ represents a division of X into discrete and continuous components, and if $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$ is the corresponding division of x, then we take $dW(x) = w(x^{\mathrm{c}}) \, dx^{\mathrm{c}} \, dV(x^{\mathrm{d}})$, where $dV(x^{\mathrm{d}})$ denotes the infinitesimal element of the Dirac delta measure which places unit mass at each atom of x^{d}. In this notation, (2.1) can be be written equivalently as

$$\text{ISE} = \sum_{x^{\mathrm{d}}} \int \{\hat{g}(y|x) - g(y|x)\}^2 \, m(x) \, w(x^{\mathrm{c}}) \, dx^{\mathrm{c}} \, dy \,, \tag{2.2}$$

where $m(x) = m^{\mathrm{c}}(x^{\mathrm{c}}|x^{\mathrm{d}}) \, P(X^{\mathrm{d}} = x^{\mathrm{d}})$, $m^{\mathrm{c}}(x^{\mathrm{c}}|x^{\mathrm{d}})$ denotes the density of X^{c} given that $X^{\mathrm{d}} = x^{\mathrm{d}}$, and the sum $\sum_{x^{\mathrm{d}}}$ is taken over all atoms of the distribution of X^{d}.

We shall assume that X^{c} and X^{d} are p- and q-variate, respectively. In practice, to overcome the "curse of dimensionality," it may be appropriate to reduce either or both of p and q. Standard dimension reduction methods can be modified for this purpose; see, for example, the methods discussed by Friedman and Stuetzle (1981), Friedman, Stuetzle and Schroeder (1984), Huber (1985), Powell, Stock and Stoker (1989) and Klein and Spady (1993). However, it should be remembered that in such cases the dimensions in which actual information is carried may not be strictly less than the dimension of the data, and that in consequence, our theoretical results in section 3 will not strictly apply after dimension reduction.

Our estimators of f and m will be of kernel type:

$$\hat{f}(x,y) = n^{-1} \sum_{i=1}^{n} K(x, X_i)\, L(y, Y_i)\,, \quad \widehat{m}(x) = n^{-1} \sum_{i=1}^{n} K(x, X_i)\,, \tag{2.3}$$

where K and L are nonnegative, generalised kernels. As (2.3) suggests, we use the same vector of smoothing parameters (one for each component) when treating the explanatory variables X_i, regardless of whether we are addressing the numerator or the denominator of the estimator $\hat{g}(y|x) = \hat{f}(x,y)/\widehat{m}(x)$.

This "convention" guarantees that for each fixed x such that $\widehat{m}(x) \neq 0$, $\hat{g}(\,\cdot\,|x)$ is a proper probability density. It also ensures that, except when $\widehat{m} = 0$, $\hat{g}$ is well defined and bounded by $\sup_{u,v} L(u,v)$. If $\widehat{m} = 0$ then $\hat{g}$ has the form $0/0$, and, for the sake of theoretical completeness, might be defined to equal an arbitrary but fixed constant. Using the same bandwidth in the numerator and denominator of $\hat{g}$ does not adversely affect the rate of convergence of estimators of g.

Next we define $K(x, X_i)$. Reflecting the division $X = (X^{\rm c}, X^{\rm d})$, write $X_i = (X_i^{\rm c}, X_i^{\rm d})$, where $X_i^{\rm d} = (X_{i1}^{\rm d}, \ldots, X_{iq}^{\rm d})$ and $X_i^{\rm c} = (X_{i1}^{\rm c}, \ldots, X_{ip}^{\rm c})$ denote the discrete and continuous components, respectively, of X_i. (In particular, we no longer use the notation $X_j^{\rm c}$ and $X_j^{\rm d}$ for the jth components of $X^{\rm c}$ and $X^{\rm d}$, respectively.) We assume $X_{ij}^{\rm d}$ takes the values $0, 1, \ldots, r_j - 1$. Put $x^{\rm c} = (x_1^{\rm c}, \ldots, x_p^{\rm c})$ and $x^{\rm d} = (x_1^{\rm d}, \ldots, x_q^{\rm d})$, and define

$$K^{\rm c}(x^{\rm c}, X_i) = \prod_{j=1}^{p} \frac{1}{h_j}\, K\biggl(\frac{x_j^{\rm c} - X_{ij}^{\rm c}}{h_j}\biggr)\,,$$

where K is a traditional kernel function (that is, a symmetric, univariate probability density), and

$$K^{\rm d}(x^{\rm d}, X_i) = \prod_{j=1}^{q} \{\lambda_j/(r_j - 1)\}^{N_{ij}(x)}\, (1 - \lambda_j)^{1 - N_{ij}(x)}\,, \tag{2.4}$$

where $N_{ij}(x) = I(X_{ij} \neq x_j^{\rm d})$, depending on $x_j^{\rm d}$ alone, and I is the usual indicator function. For $j = 1, ..., r$, the discrete kernel puts weight $1 - \lambda_j$ on $X_j^d = x_j^d$, and $\lambda_j/(r_j - 1)$ on all other $X_j^d \neq x_j^d$. In these formulae, $h_1, \ldots, h_p$ are bandwidths for the continuous components of X and satisfy $0 < h_j < \infty$, whereas $\lambda_1, \ldots, \lambda_q$ are smoothing parameters for the discrete components and are constrained by $0 \leq \lambda_j \leq (r_j - 1)/r_j$. Note that when $\lambda_j = (r_j - 1)/r_j$ assumes its upper extreme value, $K^{\rm d}(x^{\rm d}, X_i)$ becomes unrelated to $(x_j^{\rm d}, X_{ij})$ (the jth component of $x^{\rm d}$ is completely smoothed out).

Formula (2.4) describes kernels that are appropriate for unordered categorical data; see, for example, Aitchison and Aitken (1976). In the ordered case, alternative approaches can be used, employing in effect near-neighbour weights. See, for example, Wang and van Ryzin (1981), Burman (1987) and Hall and Titterington (1987). In each case the kernel weights are intrinsically different from their continuum counterparts. In particular, for the weights defined at (2.4), in the asymptotic limit as each λ_j converges to zero, $K^{\rm d}(x^{\rm d}, X_i)$ converges to 1 if $X_i = x^{\rm d}$, and converges to zero otherwise. The resulting kernel-weighted estimator of the probability at $x^{\rm d}$, converges to the naive cell-proportion, or maximum likelihood, estimator, which equals the proportion of the data for which $X_i = x^{\rm d}$.

The generalised kernels, $K(x, X_i)$ and $L(y, Y_i)$, are given by

$$K(x, X_i) = K^{\rm c}(x^{\rm c}, X_i)\, K^{\rm d}(x^{\rm d}, X_i)\,, \quad L(y, Y_i) = \frac{1}{h}\, L\biggl(\frac{y - Y_i}{h}\biggr)\,, \tag{2.5}$$

where L is another univariate kernel, possibly identical to K, and h is another bandwidth. The quantities at (2.5) are substituted into (2.3) to give $\hat f$ and $\widehat m$.

Expanding the right-hand side of (2.1) we deduce that

$$\text{ISE} = I_{1n} - 2\, I_{2n} + I_{3n}\,, \tag{2.6}$$

where

$$I_{1n} = \int \hat g(y|x)^2\, m(x)\, dW(x)\, dy\,, \quad I_{2n} = \int \hat g(y|x)\, f(x, y)\, dW(x)\, dy\,,$$

and I_{3n} does not depend on the smoothing parameters used to compute $\hat f$ and $\widehat m$. Observe that

$$I_{1n} = \int \widehat G(x)\, \frac{m(x)}{\widehat m(x)^2}\, dW(x)\,,$$

where $\widehat G(x) = \int \hat f(x, y)^2\, dy$ is expressible as

$$\widehat G(x) = \frac{1}{n^2} \sum_{i_1=1}^n \sum_{i_2=1}^n K(x, X_{i_1})\, K(x, X_{i_2}) \int L(y, Y_{i_1})\, L(y, Y_{i_2})\, dy\,.$$

Thus, the following cross-validation approximations, $\hat I_{1n}$ and $\hat I_{2n}$, to I_{1n} and I_{2n}, respectively, are motivated:

$$\hat I_{1n} = \frac{1}{n} \sum_{i=1}^n \frac{\widehat G_{-i}(X_i)\, w(X_i^{\rm c})}{\widehat m_{-i}(X_i)^2}\,, \quad \hat I_{2n} = \frac{1}{n} \sum_{i=1}^n \frac{\hat f_{-i}(X_i, Y_i)\, w(X_i^{\rm c})}{\widehat m_{-i}(X_i)}\,,$$

where the subscript $-i$ on a function of the data indicates that that quantity is computed not from the n-sample $\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ but from the $(n-1)$-sample $\mathcal{Z} \backslash \{(X_i, Y_i)\}$.

Each $\hat{I}_{jn}$ is a function of the smoothing parameters, although we have suppressed this dependence. The cross-validation criterion, CV, consists of the first two terms on the right-hand side of formula (2.6), but replaced by the above approximations:

$$\begin{aligned}\mathrm{CV}(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) = \hat{I}_{1n}(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) \\ - 2\, \hat{I}_{2n}(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)\,.\end{aligned}$$

In numerical work for cross-validation for density estimation, one generally tries to guard against using too small a value of bandwidth. If there are two or more local minima of the cross-validation criterion, one uses the second smallest of these turning points, not the smallest. Therefore, one searches up to a large positive value of bandwidth, and takes the local minimum in that range, if there is only one of these values, or the second smallest local minimum, if there is more than one. Occasionally there is no local minimum in the range, and then one takes the value at the end of the range to be the empirical bandwidth approximation.

3. MEAN SQUARED ERROR PROPERTIES

3.1. Main results. Here we describe smoothing parameters which, in asymptotic terms, are optimal for minimising the mean integrated squared error defined by taking the expected value at (2.2):

$$\begin{aligned}&\mathrm{MISE}(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) \\ &\qquad = \sum_{x^{\mathrm{d}}} \int E\{\hat{g}(y|x) - g(y|x)\}^2\, m(x)\, w(x^{\mathrm{c}})\, dx^{\mathrm{c}}\, dy\,. \qquad (3.1)\end{aligned}$$

In this formula we interpret $\hat{g}(y|x)$ as an arbitrary constant when it equals $0/0$.

Recall that $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$, where $x^{\mathrm{c}} = (x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})$ and $x^{\mathrm{d}} = (x_1^{\mathrm{d}}, \ldots, x_q^{\mathrm{d}})$, and that w is a function of x^{c}. Let $\mathcal{S}^{\mathrm{c}} = \operatorname{supp} w$ denote the support of the function w, and let $\mathcal{S}^{\mathrm{d}}$ be the support of the distribution of X^{d}. We shall assume that:

> the densities f and m have two continuous derivatives as functions of x^{c}; w is continuous, nonnegative and has compact support; $m(x)$ is bounded away from zero for $x = (x^{\mathrm{c}}, x^{\mathrm{d}}) \in \mathcal{S}^{\mathrm{c}} \times \mathcal{S}^{\mathrm{d}}$; and $\sup_{x \in \mathcal{S}^{\mathrm{c}} \times \mathcal{S}^{\mathrm{d}}} f(x, y)$ vanishes outside a compact set of values y. (3.2)

Let $f_{00}(x^{\rm c}, x^{\rm d}, y)$ $[f_{jj}(x^{\rm c}, x^{\rm d}, y)]$ denote the second derivative of $f(x^{\rm c}, x^{\rm d}, y)$ with respect to y [respectively, $x_j^{\rm c}$]. Put $\kappa = \int K^2$, $\kappa_L = \int L^2$, $\kappa_2 = \int u^2\, K(u)\, du$ and $\kappa_{L2} = \int u^2\, L(u)\, du$. Define an indicator function $I_j(u^{\rm d}, x^{\rm d})$ by

$$I_j(u^{\rm d}, x^{\rm d}) = I(u_j^{\rm d} \neq x_j^{\rm d}) \prod_{s \neq j}^{q} I(u_s^{\rm d} = x_s^{\rm d}).$$

Note that $I_j(u^{\rm d}, x^{\rm d}) = 1$ if and only if $u^{\rm d}$ and $x^{\rm d}$ differ only at the jth component. Let $a_0, \ldots, a_p$ and $b_1, \ldots, b_q$ denote real numbers. We define below a function of these quantities, which will represent the dominant term in an expansion of MISE; see (3.8) and (3.12):

$$\begin{aligned}
&\chi(a_0, \ldots, a_p, b_1, \ldots, b_q) \\
&\quad = \sum_{x^{\rm d}} \int \Bigg(\Bigg[\sum_{j=1}^{q} \frac{b_j}{r_j - 1} \sum_{u^{\rm d}} I_j(u^{\rm d}, x_j^{\rm d}) \left\{ f\big(x^{\rm c}, u^{\rm d}, y\big) - \frac{m(x^{\rm c}, u^{\rm d})}{m(x)}\, f(x, y) \right\} \\
&\qquad + \tfrac{1}{2}\, \kappa_{L2}\, a_0^2\, f_{00}(x, y) + \tfrac{1}{2}\, \kappa_2 \sum_{j=1}^{p} a_j^2 \left\{ f_{jj}(x, y) - \frac{m_{jj}(x)}{m(x)}\, f(x, y) \right\} \Bigg]^2 \\
&\qquad + \frac{\kappa^p\, \kappa_L\, f(x, y)}{a_0 \ldots a_p} \Bigg) \frac{w(x^{\rm c})}{m(x)}\, dx^{\rm c}\, dy\,, \qquad (3.3)
\end{aligned}$$

where, for $v = u$ or x, $\sum_{v^{\rm d}}$ denotes summation over atoms $v^{\rm d} = (v_1^{\rm d}, \ldots, v_q^{\rm d})$ of the distribution of $X^{\rm d}$.

Write $a_0^0, \ldots, a_p^0, b_1^0, \ldots, b_q^0$ for the values that minimise χ, subject to each of them being nonnegative. It is possible for a_j^0 or b_j^0 to be infinite. Now, $a_j^0 = 0$, for some j, only if at least one of the other a_j^0's is infinite. For the time being we exclude these degenerate cases, asking that

$$\text{the } a_j^0\text{'s and } b_j^0\text{'s are uniquely defined, and each is finite.} \qquad (3.4)$$

Therefore $0 < a_j^0 < \infty$ for each j, but it is nevertheless possible for one or more of the b_j's to vanish. The following general result may be proved. Consider the following function of positive quantities $z_0, \ldots, z_p$ and general variables $z_{p+1}, \ldots, z_{p+q}$:

$$\begin{aligned}
\chi(z_0, \ldots, z_{p+q}) &= \int \Bigg\{ \sum_{j=0}^{p+q} B_j(x, y)\, z_j \Bigg\}^2 dx\, dy + \frac{c_0}{(z_0 \ldots z_p)^{1/2}} \\
&= z^{\rm T} A z + \frac{c_0}{(z_0 \ldots z_p)^{1/2}}\,,
\end{aligned}$$

where $z = (z_0, \ldots, z_{p+q})^{\mathrm{T}}$, A is a $(p+q+1) \times (p+q+1)$ matrix, and $c_0 > 0$ is a positive constant. Then, if A is positive definite, $\chi(z_0, \ldots, z_{p+q})$ has a unique minimum, at a point where $z_0, \ldots, z_p$ are positive and finite and $z_{p+1}, \ldots, z_{p+q}$ are nonnegative and finite.

When searching for a minimum of MISE, over values of its $(p+q+1)$-variate argument, we shall confine attention to

$$\begin{aligned} &(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) \in [0, \eta]^{p+q+1}, \text{ where } \eta = \eta_n \text{ denotes} \\ &\text{any positive sequence which satisfies } n^{\epsilon} \eta_n \to \infty \text{ for each } \epsilon > 0. \end{aligned} \tag{3.5}$$

This avoids the need to treat issues addressed by Sain (2001), who pointed out that even in univariate density estimation the asymptotically optimal bandwidth, in the sense of minimising mean squared error, need not converge to zero.

Theorem 3.1. *Assume* (3.2) *and* (3.4)*, and that the smoothing parameters* h^0, $h_1^0, \ldots, h_p^0, \lambda_1^0, \ldots, \lambda_q^0$ *that minimise* MISE *are constrained by* (3.5)*. Then,*

$$\begin{aligned} h^0 \sim a_0^0 \, n^{-1/(p+5)}, \quad h_j^0 \sim a_j^0 \, n^{-1/(p+5)} \quad &\text{for} \quad 1 \le j \le p, \\ \lambda_j^0 = b_j^0 \, n^{-2/(p+5)} + o(n^{-2/(p+5)}) \quad &\text{for} \quad 1 \le j \le q, \end{aligned} \tag{3.6}$$

and $\inf \mathrm{MISE} \sim n^{-4/(p+5)} \inf \chi$.

A key assumption in Theorem 3.1 is (3.4), which excludes some cases where components of X^{d} or X^{c} contain no effective information about Y. To appreciate this point, let $Z^{\mathrm{c},-j}$ [$Z^{\mathrm{d},-j}$] denote the $(p+q)$-vector that arises after removing the jth component, X_j^{c} [respectively, X_j^{d}] of X^{c} [respectively, of X^{d}] from $Z = (X, Y)$. If X_j^{c} and $Z^{\mathrm{c},-j}$ were independent random variables, or if X_j^{d} and $Z^{\mathrm{d},-j}$ were independent, then when constructing $\hat{g} = \hat{f}/\widehat{m}$ it would make little sense to compute either $\hat{f}$ or $\widehat{m}$ using the full data vectors (X_i, Y_i). We would instead delete the jth component of the continuous part of X_i, or of the discrete part of X_i, respectively.

In the first of these cases, $f_{jj} = (m_{jj}/m)\, f$, and so the term in a_j^2 vanishes from the right-hand side of (3.3). As a result, $a_j^0 = \infty$, and so (3.4) is violated. It can be restored by excluding the jth component of X^{c}, as argued above. In the second case the quantity b_j in (3.3) can be absorbed into the other b_k's, and so the minimiser of χ is not uniquely defined. Therefore, (3.4) again fails, but can be restored by dropping the jth component of X^{d}. In these respective instances it is fair to say that X_j^{c}, or X_j^{d}, is "completely irrelevant" for estimating g.

If $X_j^{\rm d}$ were independent of Y, but not independent of the other components of X, then although in some respects the correct approach would be to delete the jth component of the discrete part of each data vector, failing to do so would not necessarily mean that (3.4) was violated. This reflects the fact that the full data vectors can produce estimators of g with lower mean squared error, and so may be beneficial. A similar argument applies if $X_j^{\rm c}$ is independent of Y but not of other components of X, although in this case a relatively standard kernel approach to estimation, advocated at (2.3), is not the best way to proceed.

3.2. Proof of Theorem 3.1. Using (3.7) below it may be shown that ISE does not converge to zero unless $Vh \to \infty$ as $n \to \infty$, where $V = nh_1 \ldots h_p$. Therefore we may, without loss of generality, add to (3.5) the constraint that $Vh \geq t_n$, where $\{t_n\}$ is an unspecified sequence of constants diverging to infinity.

Note that $\hat g = (g + \delta_f)/(1 + \delta_m)$, where $\delta_f = (\hat f - f)/m$ and $\delta_m = (\widehat m - m)/m$. By Taylor expansion,

$$\hat g(y|x) - g(y|x) = \left\{\hat f(x,y) - \widehat m(x)\, g(y|x) + Q(x,y)\right\} m(x)^{-1}\,, \tag{3.7}$$

where $Q = -m\delta_m\delta_f + (\delta_m^2 - \delta_m^3 + \ldots)\,(f + m\delta_f)$ consists of quadratic and higher-order terms in δ_f and δ_m. Using these expansions and the methods we shall employ below to approximate

$$\mathrm{MISE}_1 = \sum_{x^{\rm d}} \int E\{\hat f(y|x) - \widehat m(x)\, g(y|x)\}^2\, \frac{w(x^{\rm c})}{m(x)}\, dx^{\rm c}\, dy\,,$$

and noting the convention that $\hat g$ is taken to equal a constant if it would otherwise equal $0/0$, it may be proved that

$$\mathrm{MISE} = \mathrm{MISE}_1 + o(\eta_1)\,, \tag{3.8}$$

uniformly in smoothing-parameter vectors in $[0, \eta]^{p+q+1}$ satisfying $Vh \geq t_n$, where $\eta_1 = \eta_2 + \eta_3$, $\eta_2 = \sum_j \lambda_j + \sum_j h_j^2 + h^2$ and $\eta_3 = (Vh)^{-1}$.

Put $\rho_j = \lambda_j/\{(1 - \lambda_j)\,(r_j - 1)\}$, and let $\psi(x^{\rm c}, y \,|\, x^{\rm d})$ denote the density of $(X^{\rm c}, Y)$ given $X^{\rm d}$. Write $\psi_{00}(x^{\rm c}, y \,|\, x^{\rm d})$ $[\psi_{jj}(x^{\rm c}, y \,|\, x^{\rm d})]$ for the second derivative of $\psi(x^{\rm c}, y \,|\, x^{\rm d})$ with respect to y [respectively, $x_j^{\rm c}$]. Given members $x^{\rm d} = (x_1^{\rm d}, \ldots, x_q^{\rm d})$ and $u^{\rm d} = (u_1^{\rm d}, \ldots, u_q^{\rm d})$ of the sample space of $X^{\rm d}$, let $u_j^{\rm d}(x) = I(u_j^{\rm d} \neq x_j^{\rm d})$. In this notation,

$$E\{\hat f(x,y)\} = \sum_{u^{\rm d}} P\big(X^{\rm d} = u^{\rm d}\big) \left\{\prod_{j=1}^q (1 - \lambda_j)\, \rho_j^{u_j^{\rm d}(x)}\right\} \int \left\{\prod_{j=1}^p K(z_j)\right\} L(v)$$

$$\times\,\psi\Big(x_1^{\rm c}-h_1z_1,\ldots,x_p^{\rm c}-h_pz_p,y-hv\,\Big|\,u_1^{\rm d},\ldots,u_q^{\rm d}\Big)\,dz_1\ldots dz_p\,dv$$
$$=f(x,y)+\sum_{j=1}^q\,\frac{\lambda_j}{r_j-1}\Big\{\sum_{u^{\rm d}}\,I_j(u^{\rm d},x^{\rm d})\,f\big(x^{\rm c},u^{\rm d},y\big)-f(x,y)\Big\}$$
$$+\,\tfrac12\,\kappa_{L2}\,h^2\,f_{00}(x,y)+\tfrac12\,\kappa_2\sum_{j=1}^p\,h_j^2\,f_{jj}(x,y)+o(\eta_2)\,,$$

where the remainders here and in (3.9)–(3.11) below are of the stated size uniformly in $x^{\rm c}\in\operatorname{supp}w$, in $x^{\rm d}$ in the support of the distribution of $X^{\rm d}$, and in y, as well as in smoothing-parameter vectors in $[0,\eta]^{p+q+1}$ satisfying $Vh\ge t_n$.

Similarly,

$$E\{\widehat m(x)\}=m(x)+\sum_{j=1}^q\,\frac{\lambda_j}{r_j-1}\,\Big\{\sum_{u^{\rm d}}\,I_j(u^{\rm d},x^{\rm d})\,m\big(x^{\rm c},u^{\rm d}\big)-m(x)\Big\}$$
$$+\,\tfrac12\,\kappa_2\sum_{j=1}^p\,h_j^2\,m_{jj}(x)+o(\eta_2)\,. \tag{3.9}$$

Therefore,

$$E\{\hat f(x,y)\}-E\{\widehat m(x)\}\,g(y|x)$$
$$=\sum_{j=1}^q\,\frac{\lambda_j}{r_j-1}\,\sum_{u^{\rm d}}\,I_j(u^{\rm d},x^{\rm d})\,\Big\{f\big(x^{\rm c},u^{\rm d},y\big)-\frac{m(x^{\rm c},u^{\rm d})}{m(x)}\,f(x,y)\Big\}$$
$$+\,\tfrac12\,\kappa_{L2}\,h^2\,f_{00}(x,y)+\tfrac12\,\kappa_2\sum_{j=1}^p\,h_j^2\,\Big\{f_{jj}(x,y)-\frac{m_{jj}(x)}{m(x)}\,f(x,y)\Big\}+o(\eta_2)\,. \tag{3.10}$$

Note too that

$$n\,\operatorname{var}\,\{\hat f(x,y)-\widehat m(x)\,g(y|x)\}$$
$$=\operatorname{var}\,\big[K(x,X_i)\,\{L(y,Y_i)-g(y|x)\}\big]=E\{K(x,X_i)\,L(y,Y_i)\}^2+o(\eta_3)$$
$$=n\,\kappa^p\,\kappa_L\,f(x,y)\,\eta_3+o(\eta_3)\,. \tag{3.11}$$

Combining (3.10) and (3.11) we deduce that

$$\mathrm{MISE}_1(h,h_1,\ldots,h_p,\lambda_1,\ldots,\lambda_q)=n^{-4/(p+1)}\,\chi(a_0,\ldots,a_p,b_1,\ldots,b_q)+o(\eta_1)\,, \tag{3.12}$$

uniformly in smoothing-parameter vectors in $[0,\eta]^{p+q+1}$ satisfying $Vh\ge t_n$, where the scalars $a_0,\ldots,a_p,b_1,\ldots,b_q$ are defined by $h=a_0\,n^{-1/(p+5)}$, $h_j=a_j\,n^{-1/(p+5)}$ and $\lambda_j=b_j\,n^{-2/(p+5)}$. The theorem follows from (3.8) and (3.12).

4. PROPERTIES OF CROSS-VALIDATION

Recall from section 2 that $X = (X^{\rm c}, X^{\rm d})$, where $X^{\rm c}$ and $X^{\rm d}$ are p- and q-variate, respectively. We shall assume that only the first p_1 components of $X^{\rm c}$, and the first q_1 components of $X^{\rm d}$, are "relevant" to estimating the distribution of Y given X, the others being "irrelevant" in the sense defined below:

> for integers $0 \le p_1, p_2 \le p$ and $0 \le q_1, q_2 \le q$ satisfying $p_1 + p_2 = p$ and $q_1 + q_2 = q$, the following is true: the vector $X^{(1)}$ comprised of the first p_1 components of $X^{\rm c}$, and the first q_1 components of $X^{\rm d}$, is stochastically independent of the vector $X^{(2)}$ consisting of the last p_2 components of $X^{\rm c}$, and the last q_2 components of $X^{\rm d}$; and $X^{(2)}$ and Y are independent. (4.1)

A technical definition of "relevance" of the first p_1 components of $X^{\rm c}$, and first q_1 components of $X^{\rm d}$, will be given at (4.3).

For the p_1 relevant continuous components, and q_1 relevant discrete components, of X, we should impose the analogue of assumption (3.4), asserting that the asymptotically optimal smoothing parameters are of conventional size. To this end we introduce the following analogue of the function χ at (3.3), now tailored to just the relevant components of X:

$$\begin{aligned}
&\bar\chi(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\
&\quad = \sum_{\bar x^{\rm d}} \int \Bigg(\Bigg[\sum_{j=1}^{q_1} \frac{b_j}{r_j - 1} \sum_{\bar u^{\rm d}} I_j(\bar u^{\rm d}, \bar x^{\rm d}) \left\{ \bar f(\bar x^{\rm c}, \bar u^{\rm d}, y) - \frac{\bar m(\bar x^{\rm c}, \bar u^{\rm d})}{\bar m(\bar x)}\, \bar f(\bar x, y) \right\} \\
&\qquad + \tfrac{1}{2}\, \kappa_{L2}\, a_0^2\, \bar f_{00}(\bar x, y) + \tfrac{1}{2}\, \kappa_2 \sum_{j=1}^{p_1} a_j^2 \left\{ \bar f_{jj}(\bar x, y) - \frac{\bar m_{jj}(\bar x)}{\bar m(\bar x)}\, \bar f(\bar x, y) \right\} \Bigg]^2 \\
&\qquad + \frac{\kappa^{p_1}\, \kappa_L\, \bar f(\bar x, y)}{a_0 \ldots a_{p_1}} \Bigg) \bar w(\bar x^{\rm c}, \bar x^{\rm d})\, d\bar x^{\rm c}\, dy\,,
\end{aligned} \tag{4.2}$$

where

$$\bar w(\bar x^{\rm c}, \bar x^{\rm d}) = \sum_{x^{\rm d}_{q_1+1}, \ldots, x^{\rm d}_q} \int \frac{w(\bar x^{\rm c}, x^{\rm c}_{p_1+1}, \ldots, x^{\rm c}_p)}{m(\bar x^{\rm c}, x^{\rm c}_{p_1+1}, \ldots, x^{\rm c}_p, \bar x^{\rm d}, x^{\rm d}_{q_1+1}, \ldots, x^{\rm d}_q)}\, dx^{\rm c}_{p_1+1} \ldots dx^{\rm c}_p$$

and "bar" notation refers to functions or vectors involving only the first p_1 continuous components and the first q_1 discrete components. For example, $\bar x^{\rm c}$ is the vector consisting of the first p_1 components of $x^{\rm c}$, $\bar f$ denotes the joint density of the first p_1 components of $X^{\rm c}$ and of Y, $\bar f_j$ denotes the jth derivative of $\bar f$ with respect to the jth component of $\bar x^{\rm c}$, and so on.

The function $\bar{\chi}$ coincides exactly with χ in the case where the last p_2 a_j's, and last q_2 b_j's, are deleted from the argument of χ, and the weight $w(x^{\mathrm{c}})/m(x)$ at (3.3) is replaced by $\bar{w}(\bar{x}^{\mathrm{c}}, \bar{x}^{\mathrm{d}})$. Of course, $\bar{w}(\bar{x}^{\mathrm{c}}, \bar{x}^{\mathrm{d}})$ is obtained from $w(x^{\mathrm{c}})/m(x)$ on integrating (and summing) out the irrelevant components of x, and is the weight function appropriate to the mean integrated squared error that would be obtained at (3.1) if we were to drop all irrelevant components from the estimator $\hat{g}(y|x)$ appearing there. Therefore, we expect the optimal values of smoothing parameters in the present problem to be exactly those given by Theorem 3.1, but with (p, q) at (3.6) replaced by (p_1, q_1), and $h^0, h_1^0, \ldots, h_{p_1}^0, \lambda_1^0, \ldots, \lambda_{q_1}^0$ there chosen to minimise $\bar{\chi}$, defined at (4.2), rather than χ, given at (3.3). Theorem 4.1 below will show that cross-validation selects smoothing parameters for the relevant components of X in precisely this asymptotically optimal manner.

Write $a_0^0, \ldots, a_{p_1}^0, b_1^0, \ldots, b_{q_1}^0$ for the values that minimise $\bar{\chi}$, subject to each of them being nonnegative. The analogue of condition (3.4) is:

$$\text{the } a_j^0\text{'s and } b_j^0\text{'s are uniquely defined, and each is finite.} \tag{4.3}$$

In order to be able to detect the effect of relevant components of X on the conditional distribution of Y, within the domain to which we are constrained by the weight function w, we assume that

$$\sup_{\bar{x} \in \mathrm{supp}\, \bar{w}} \sup_y \bar{g}(y \,|\, \bar{x}) > 0\,. \tag{4.4}$$

The empirical observation that smoothing parameters chosen by cross-validation diverge to their upper extremes when the respective components of X are irrelevant, reflects the fact that cross-validation attempts to shrink the distributions of irrelevant components to the least variable, uniform case. There they have least impact on the variance terms of curve estimators; the fact that they contain no information about Y means that they have little impact on bias. However, if the irrelevant components of X are already uniformly distributed then the effect of choosing the respective smoothing parameters may be comparatively small, and so different behaviour can be expected. For the sake of simplicity we shall regard the case of uniformly distributed irrelevant components as pathological, and impose a regularity condition to eliminate it, as follows.

Define a kernel ratio for irrelevant data components:

$$R(\bar{x}, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q)$$

$$= \frac{E\Big[\Big\{\prod_{j=p_1+1}^{p} K\Big(\frac{\bar{x}_j^{\rm c} - X_{1j}^{\rm c}}{h_j}\Big)^2\Big\} \prod_{j=q_1+1}^{q} \big\{(1-\lambda_j)\,\rho_j^{N_{1j}(\bar{x})}\big\}^2\Big]}{\Big(E\Big[\Big\{\prod_{j=p_1+1}^{p} K\Big(\frac{\bar{x}_j^{\rm c} - X_{1j}^{\rm c}}{h_j}\Big)\Big\} \prod_{j=q_1+1}^{q} \big\{(1-\lambda_j)\,\rho_j^{N_{1j}(\bar{x})}\big\}\Big]\Big)^2} \,.$$

Note that, by Hölder's inequality, $R \geq 1$ for all choices of $h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q$. It is easy to see that, provided $K(0) > K(\delta)$ for all $\delta > 0$, $R \to 1$ as $h_j \to \infty$ (for $p_1 + 1 \leq j \leq p$) and $\lambda_j \to (r_j - 1)/r_j$ (for $q_1 + 1 \leq j \leq q$). Generally speaking, however, $R > 1$ for other values of these smoothing parameters. Exceptions to this rule can arise if marginal distributions are uniform, however. We shall eliminate problems of this type by assuming that:

the only values of $h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q$, in the range $h_j \geq 0$ and $0 \leq \lambda_j \leq (r_j - 1)/r_j$, for which $R(\bar{x}, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q) = 1$ for some $\bar{x} \in \operatorname{supp} \bar{w}$, are $h_j = \infty$ for $p_1 + 1 \leq j \leq p$ and $\lambda_j = (r_j - 1)/r_j$ for $q_1 + 1 \leq j \leq q$. (4.5)

As discussed in section 3, we need to ensure that the cross-validation algorithm is not seduced by the possibility that there exist nonvanishing smoothing parameters that produce estimators with zero bias. In section 3 we averted this problem by treating only smoothing parameters that converge to 0; see assumption (3.5). Here that is not desirable, however, since we expect that smoothing parameters corresponding to irrelevant components will be bounded away from zero. Constraint (3.5) would have us assume in advance that the bandwidths associated with relevant components converge to zero, yet for the sake of realism we do not wish those components to be identified to the experimenter. Therefore we shall take a different tack, as follows.

Note that, in view of (4.1), the contributions of irrelevant components cancel from the ratio $\bar{\mu}_g(y|\bar{x}) = E\{\hat{f}(x,y)\}/E\{\widehat{m}(x)\}$. Let $\bar{g}$ denote the version of g when irrelevant components are dropped. We shall assume that:

$\int dy \int_{\operatorname{supp} \bar{w}} \{\bar{\mu}_g(y|\bar{x}) - \bar{g}(y|\bar{x})\}^2 \, d\bar{x}$, interpreted as a function of $h_1, \ldots, h_{p_1}$ and $\lambda_1, \ldots, \lambda_{q_1}$, vanishes if and only if all those smoothing parameters vanish. (4.6)

Finally we suppose conventional conditions on the bandwidths and kernels. Define

$$H = H(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) = h \Big(\prod_{j=1}^{p_1} h_j\Big) \prod_{j=p_1+1}^{p_2} \min(h_j, 1) \,,$$

let $\Lambda_j = [0, (r_j - 1)/r_j]$ denote the range of possible values taken by λ_j, and let $0 < \epsilon < 1/(p+5)$. Assume that:

$$\begin{array}{l} 0 < h \le n^{-\epsilon};\ n^{\epsilon-1} \le H \le n^{-\epsilon};\ h_1 \dots h_{p_1} h \ge n^{\epsilon-1};\ \text{the} \\ \text{kernels } K \text{ and } L \text{ are symmetric, compactly supported,} \\ \text{Hölder-continuous probability densities; and } K(0) \ne 0. \end{array} \tag{4.7}$$

We also impose the obvious requirement that the bandwidths that minimise CV are chosen so that CV is well-defined, which in particular means that the estimator $\widehat{m}$ should not vanish on $\operatorname{supp} w$. This implies that for some $C_1 > 0$, the probability none of the bandwidths $h, h_1, \dots, h_p$ is less than n^{-C_1} converges to 1 as $n \to \infty$. In company with (4.7), this entails that for some $C_2 > 0$, the probability that none of the bandwidths $h, h_1, \dots, h_{p_1}$ exceeds n^{C_2} converges to 1 as $n \to \infty$. Therefore, it may be supposed without loss of generality that

$$\text{for some } C > 0,\ \min(h, h_1, \dots, h_p) > n^{-C} \text{ and } \max(h_1, \dots, h_p) \le n^C. \tag{4.8}$$

This property will be used in several places in our proofs, for example in the second-last paragraph of step (i), although we do not list it among the regularity conditions in Theorem 4.1.

We should comment on the extent to which (4.7) and (4.8) require knowledge of p_1, the number of relevant components of X. The conditions do betray knowledge of whether p_1 is zero or positive. However, we claim that under the following side condition, which does not depend on p_1, they hold whenever $1 \le p_1 \le p$:

$$\begin{array}{l} \text{no } h_j, \text{ for } 1 \le j \le p, \text{ takes a value exceeding } D \log n, \text{ where } D > 0 \\ \text{is arbitrary; } \prod_{j=1}^p h_j \in [n^{\delta-1}, n^{-\delta}] \text{ for some } \delta \in (0, \tfrac{1}{2}); \text{ and } 1 \le p_1 \le p. \end{array} \tag{4.9}$$

If this constraint holds then (4.7) and (4.8) obtain for any $\epsilon \in (0, \delta)$ and $C \ge 1$, and all sufficiently large n, regardless of the value of p_1. Condition (4.9) is of course reasonable, in that it does not prevent $h_1, \dots, h_{p_1}$ from taking appropriately small values, and at the same time it allows $h_{p_1+1}, \dots, h_p$ to assume values that diverge to infinity.

Theorem 4.1. *Assume* (3.2), (4.1) *and* (4.4)–(4.7), *and let* $\hat{h}, \hat{h}_1, \dots, \hat{h}_p, \hat{\lambda}_1, \dots, \hat{\lambda}_q$ *denote the smoothing parameters that minimise* CV *subject to the bandwidth constraints imposed in* (4.7). *Then, interpreting each convergence of a sequence of*

random variables as convergence in probability, we have:

$$
\begin{aligned}
& n^{1/(p_1+5)}\,\hat{h}^0 \to a_0^0\,, \quad n^{1/(p_1+5)}\,\hat{h}_j^0 \to a_j^0 \quad \textit{for} \quad 1 \le j \le p_1\,, \\
& P(\hat{h}_j > C) \to 1 \quad \textit{for} \quad p_1+1 \le j \le p \quad \textit{and all} \quad C > 0\,, \\
& n^{2/(p_1+5)}\,\hat{\lambda}_j^0 \to b_j^0 \quad \textit{for} \quad 1 \le j \le q_1\,, \\
& \hat{\lambda}_j \to (r_j-1)/r_j \quad \textit{for} \quad q_1+1 \le j \le q\,,
\end{aligned}
$$

and $n^{4/(p_1+5)} \inf \mathrm{MISE} \to \inf \chi$.

The conclusions of Theorem 4.1 may be summarised as follows. The smoothing parameters chosen by cross-validation, and corresponding to relevant components of the variables X_i, have the properties of asymptotic optimality described by Theorem 3.1. On the other hand, the cross-validation smoothing parameters that correspond to irrelevant components converge in probability to the upper extremities of their respective ranges.

It is always possible, in practice, that the method makes a mistake and, in effect, incorrectly removes relevant variables by choosing a too-large value of bandwidth. Results such as Theorem 4.1 state that the probability that this happens converges to 1 as $n \to \infty$, but nevertheless there is always a nonzero probability that the method does the wrong thing in finite-sample settings.

Next we discuss the performance of the empirical smoothing parameters when they are used to construct $\hat{g}$ at a point. We shall show that they produce an estimator that has the same first-order properties it would enjoy if the asymptotically optimal, deterministic parameters were employed. The latter may be defined thus: $h = n^{-1/(p_1+5)}\,a_0^0$, $h_j = n^{-1/(p_1+5)}\,a_j^0$ for $1 \le j \le p_1$, $\lambda_j = n^{-2/(p_1+5)}\,b_j^0$ for $1 \le j \le q_1$, $h_j \to \infty$ for $p_1+1 \le j \le p$, and $\lambda_j \to (r_j-1)/r_j$ for $q_1+1 \le j \le q$, where $a_0^0, a_1^0, \ldots, a_{p_1}^0, b_1^0, \ldots, b_{q_1}^0$ minimise $\bar{\chi}$ defined at (4.2).

If $\hat{g}$ is computed using the asymptotically optimal, deterministic smoothing parameters, then

$$
\hat{g}(y|x) = g(y|x) + n^{-2/(p_1+5)} \left\{ \beta(\bar{x}, y) + \sigma(\bar{x}, y)\, Z_n(x, y) \right\} + o_p\big(n^{-2/(p_1+5)}\big)\,, \quad (4.10)
$$

where the random variable $Z_n(x, y)$ has the standard normal distribution,

$$
\beta(\bar{x}, y) = \sum_{j=1}^{q_1} \frac{b_j}{r_j - 1} \sum_{\bar{u}^{\mathrm{d}}} I_j(\bar{u}^{\mathrm{d}}, \bar{x}^{\mathrm{d}}) \left\{ \bar{g}\big(y \,\big|\, \bar{x}^{\mathrm{c}}, \bar{u}^{\mathrm{d}}\big) - \frac{\bar{m}(\bar{x}^{\mathrm{c}}, \bar{u}^{\mathrm{d}})}{\bar{m}(\bar{x})}\, \bar{g}(y|\bar{x}) \right\}
$$

$$+ \tfrac{1}{2}\,\kappa_{L2}\,a_0^2\,\bar{g}_{00}(y|\bar{x}) + \tfrac{1}{2}\,\kappa_2 \sum_{j=1}^{p_1} a_j^2 \left\{ \bar{g}_{jj}(y|\bar{x}) - \frac{\bar{m}_{jj}(\bar{x})}{\bar{m}(\bar{x})}\,\bar{g}(y|\bar{x}) \right\},$$

$$\sigma(\bar{x}, y)^2 = \frac{\kappa^{p_1}\,\kappa_L\,\bar{g}(y|\bar{x})}{a_0 \ldots a_{p_1}}$$

denote asymptotic bias and variance respectively, and $\bar{g}_{jj}(y|\bar{x})$ is the second derivative of $\bar{g}(y|\bar{x})$ with respect to $\bar{x}_j^{\rm c}$. We shall give a proof of (4.10) as part of our derivation of Theorem 4.2 in section 7. The theorem argues that (4.10) continues to hold if we choose the smoothing parameters empirically, by cross-validation.

Recall that $\mathcal{S}^{\rm c} = \operatorname{supp} w$ and $\mathcal{S}^{\rm d}$ denotes the support of the distribution of $X^{\rm d}$.

Theorem 4.2. *Assume the conditions imposed in Theorem 4.1, let* $\hat{h}, \hat{h}_1, \ldots, \hat{h}_p$, $\hat{\lambda}_1, \ldots, \hat{\lambda}_q$ *denote the empirically chosen smoothing parameters prescribed there, and let* $x = (x^{\rm c}, x^{\rm d}) \in \mathcal{S}^{\rm c} \times \mathcal{S}^{\rm d}$. *Then* (4.10) *remains true, for the same functions* β *and* σ, *if* $\hat{g}(y|x)$ *is computed using the smoothing parameters chosen by cross-validation, rather than the asymptotically optimal, deterministic parameters.*

Up to now we have assumed that the dependent variable Y is continuous. If instead Y is discrete, taking r different values, it is necessary to replace $L(y, Y_i) = h^{-1}L\{(y-Y_i)/h\}$, defined at (2.5), by $L(y, Y_i) = \lambda^{N_i(y)}(1-\lambda)^{1-N_i(y)}$, where $N_i(y) = I(Y_i \neq y)$. Then Theorem 4.2 needs to be modified by replacing $p_1 + 5$ by $p_1 + 4$, and replacing $n^{1/(p_1+5)}\hat{h} \to a_0^0$ by $n^{2/(p_1+4)}\hat{\lambda} \to b^0$, where b^0 is defined in a way similar to b_j^0 for $j = 1, \ldots, q_1$.

5. NUMERICAL SIMULATION AND PRACTICAL EXAMPLES

5.1. Monte Carlo Study

In this section we outline some modest Monte Carlo experiments designed to investigate the performance of the proposed estimator. Simulation code was written in ANSI C, and simulations were run on a 2.4Ghz Pentium IV running FreeBSD 5.2 Current. Numerical bandwidth search was conducted using Powell's quadratically convergent direction set method (Numerical Recipes in C), and the search algorithm was restarted twice from different random starting values with those bandwidths yielding the lowest value of the objective function being retained for estimation. Search tolerances were set for an absolute tolerance of 1.49012×10^{-8} and relative tolerance of 1.19209×10^{-7}. For the smoothing parameters associated with the continuous variables, the range of values over which search occurred was

$(0, 1.79779 \times 10^{308})$, the upper bound corresponding to the largest possible double precision number for this processor. Execution speed for the cross-validation procedure for $n = 100$ was around 20 seconds for $P = 2$, and involved roughly 1,000 objective function evaluations for convergence. However, the algorithm is of numerical order n^3, hence execution time increases rather rapidly as the sample size increases.

1. Relevant X_1, X_2, $\rho_{x_1,x_2} = 0.0$

Let $Z = (Y, X_1, X_2) = (Y, X)$, and let

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix} = \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} = \begin{pmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.0 \\ 0.5 & 0.0 & 1.0 \end{pmatrix}.$$

We generate Z to have a multivariate normal distribution having mean μ and covariance Σ, hence, the conditional distribution of Y given X is also normally distributed. We draw independent estimation and evaluation samples of size $n_1 = 50$, 100 and $n_2 = 1,000$, respectively. We consider three models, a correctly specified parametric model, an incorrectly specified parametric model (conditional exponential distribution) and the proposed nonparametric estimator. We draw 1,000 Monte Carlo replications, estimate each model, and compute their respective predicted RMSEs given by $\sqrt{n_2^{-1} \sum_{i=1}^{n_2} (\hat{f}(y_i|x_i) - f(y_i|x_i))^2}$, where $f(y_i|x_i)$ is the true conditional density function, and $\hat{f}(y_i|x_i)$ one of the estimates. We then report the median root mean square error of each of the three estimators along with a summary of the cross-validated bandwidth constants c_y and c_x's ($h_y = c_y \sigma_y n^{-1/7}$, $h_{x_j} = c_x \sigma_{x_j} n^{-1/7}$, $j = 1, 2$).

Table 1. Median estimator RMSE on independent evaluation data, relevant X_1, X_2, $\rho_{x_1,x_2} = 0.0$.

n	Kernel	Correct	Misspecified
50	0.159	0.066	0.394
100	0.134	0.046	0.395

Table 2. Median bandwidth constant c ($h = c\sigma n^{-1/7}$), upper and lower quartiles are in parentheses, relevant X_1, X_2, $\rho_{x_1,x_2} = 0.0$.

n	h_{x_1}	h_{x_2}	h_y
50	1.058	0.992	0.939
	(0.711,1.44)	(0.676,1.38)	(0.726,1.09)
100	1.056	1.097	0.916
	(0.81,1.32)	(0.838,1.36)	(0.744,1.04)

2. Relevant X_1, X_2, $\rho_{x_1,x_2} = 0.25$

Now let μ remain the same as before and let

$$\Sigma = \begin{pmatrix} 1.00 & 0.50 & 0.50 \\ 0.50 & 1.00 & 0.25 \\ 0.50 & 0.25 & 1.00 \end{pmatrix}.$$

Table 3. Median estimator RMSE on independent evaluation data, relevant X_1, X_2, $\rho_{x_1,x_2} = 0.25$.

n	Kernel	Correct	Misspecified
50	0.135	0.061	0.357
100	0.113	0.042	0.357

Table 4. Median bandwidth constant c ($h = c\sigma n^{-1/7}$), upper and lower quartiles are in parentheses, relevant X_1, X_2, $\rho_{x_1,x_2} = 0.25$.

n	h_{x_1}	h_{x_2}	h_y
50	1.246	1.241	0.933
	(0.781,1.94)	(0.826,1.86)	(0.706,1.08)
100	1.293	1.262	0.911
	(0.957,1.66)	(0.9,1.66)	(0.724,1.03)

Next, we consider the case for which X_2 is irrelevant.

3. Irrelevant X_2, $\rho_{x_1,x_2} = 0.0$

Let

$$\Sigma = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.5 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}.$$

Table 5. Median estimator RMSE on independent evaluation data, irrelevant X_2, $\rho_{x_1,x_2} = 0.0$.

n	Kernel	Correct	Misspecified
50	0.109	0.054	0.316
100	0.086	0.038	0.316

Table 6. Median bandwidth constant c ($h = c\sigma n^{-1/7}$), upper and lower quartiles are in parentheses, irrelevant X_2, $\rho_{x_1,x_2} = 0.0$.

n	h_{x_1}	h_{x_2}	h_y
50	1.120	5.781	0.951
	(0.765,1.57)	(1.53,1.25×10^7)	(0.674,1.12)
100	1.048	6.871	0.938
	(0.746,1.32)	(1.88,7.88×10^6)	(0.73,1.07)

4. Irrelevant X_2, $\rho_{x_1,x_2} = 0.25$

Finally, we let

$$\Sigma = \begin{pmatrix} 1.00 & 0.50 & 0.00 \\ 0.50 & 1.00 & 0.25 \\ 0.00 & 0.25 & 1.00 \end{pmatrix}.$$

Table 7. Median estimator RMSE on independent evaluation data, irrelevant X_2, $\rho_{x_1,x_2} = 0.25$.

n	Kernel	Correct	Misspecified
50	0.114	0.056	0.323
100	0.094	0.040	0.323

Table 8. Median bandwidth constant c ($h = c\sigma n^{-1/7}$), upper and lower quartiles are in parentheses, irrelevant X_2, $\rho_{x_1,x_2} = 0.25$.

n	h_{x_1}	h_{x_2}	h_y
50	1.092	4.220	0.955
	(0.711,1.52)	(1.42,1.19×10^7)	(0.703,1.11)
100	1.042	4.115	0.921
	(0.752,1.34)	(1.73,7.55×10^6)	(0.729,1.05)

Summarizing the above simulation results, we observe that

1) As expected, the nonparametric method is less efficient than the correctly specified parametric model, but can be much more efficient than an incorrectly specified parametric model.

2) As n_1 increases, estimation RMSE falls for the nonparametric model but not for the incorrectly specified parametric model revealing the consistency of the nonparametric approach and the inconsistency of incorrectly specified parametric models.

3) For the irrelevant X_2, cross-validation oversmooths the irrelevant variable quite dramatically. The median value of the bandwidth for irrelevant X_2 is at least 400% that for the relevant X_1.

4) For the case in which the irrelevant X_2 is correlated with the relevant X_1, cross-validation also oversmooths X_2 quite dramatically.

We have also done numerous simulations involving discrete Y and mixed discrete and continuous X. The results show that the cross-validation method has a high probability of selecting the upper bound value for λ when a discrete variable is irrelevant. The detailed results are not reported here to conserve space. In the remaining part of this section we present two empirical applications.

5.2. Firm Investment and Predictor Relevance.

Hsiao & Tahmiscioglu (1997) addressed the issue of financial constraints on company investment for a panel of US firms. The authors of that study have graciously provided us with an updated version of their data covering 432 firms for the years 1982 through 1998 which we use to examine the behavior of cross-validated bandwidth selection for conditional density estimation for a substantive current data set.

The variables contained in the dataset are the beginning of the period capital stock for the ith firm in period t, K_{it}, the ratio of capital investment to the capital stock for the ith firm in period t, I_{it}/K_{it}, a liquidity variable (defined as cash flow minus dividends), L_{it}/K_{it}, the ratio of sales to the capital stock, S_{it}/K_{it}, Tobin's q, q_{it},[1] a firm specific index $F_i = 1, \ldots, 432$, and year $T_t = 1982, \ldots, 1998$. A number of investment equations have been specified, and various regression models of the relationship

$$\left(\frac{I}{K}\right)_{it} = m\left\{\left(\frac{I}{K}\right)_{i,t-1}, \left(\frac{L}{K}\right)_{i,t-1}, \left(\frac{S}{K}\right)_{i,t-1}, q_{it}, F_i, T_t\right\} + \epsilon_{it}$$

can be found throughout the literature; see Hsiao & Tahmiscioglu (1997) and the references therein. However, it has been noted that "Estimates derived from these different approaches can be very different and hence lead to very different inferences" (Hsiao & Tahmiscioglu (1997), page 456).

We estimate the conditional density of $\left(\frac{I}{K}\right)_{it}$ given the remaining predictors. We model F_i as a nominal variable and T_t as an ordinal one, the remaining being treated as continuous. Table 9 summarizes the cross-validated bandwidths.

Table 9. Cross-validated bandwidths. The upper bound values for $\lambda \approx 1$.

Variable	Bandwidth
$\left(\frac{I}{K}\right)_{it}$	$h = 0.596\sigma n^{-1/(5+p)}$
$\left(\frac{I}{K}\right)_{i,t-1}$	$h = 1.25 \times 10^{19}\sigma n^{-1/(5+p)}$
$\left(\frac{L}{K}\right)_{i,t-1}$	$h = 5.73\sigma n^{-1/(5+p)}$
$\left(\frac{S}{K}\right)_{i,t-1}$	$h = 6.24 \times 10^{15}\sigma n^{-1/(5+p)}$
q_{it}	$h = 0.408\sigma n^{-1/(5+p)}$
F_i	$\lambda = 0.232$
T_t	$\lambda = 0.257$

It can be seen from Table 9 that lagged investment and lagged sales are smoothed out from the resulting conditional density suggesting that they have a uniform marginal distribution with respect to current investment. However, we also observe that liquidity is not smoothed out, in agreement with the findings of

[1] A ratio devised by Tobin who hypothesized that the combined market value of all the companies on the stock market should be about equal to their replacement costs. The individual calculation is the market value of a firm's assets divided by their replacement value.

Hsiao & Tahmiscioglu (1997) who conclude that "liquidity is an important determinant of investment" (page 456), a finding that has "been largely dismissed by economists following the neoclassical tradition" (pg 455).

5.3. Veterans Lung Cancer Data.

We consider the dataset found in Kalbfleisch and Prentice (1980, pp. 223–224), which models survival in days of cancer patients using six categorical explanatory variables: treatment type, cell type, Karnofsky score, months from diagnosis, age in years, and prior therapy. The dataset contains 137 observations, and the number of cells greatly exceeds the number of observations. Clearly, the conventional frequency nonparametric method cannot be used for this dataset. We create a binary outcome taking values 1 if survival is less than or equal to 180 days, 0 otherwise, and consider the performance of the proposed estimator versus a parametric Probit estimator. We wish to evaluate the true predictive performance of each estimator. To this end we randomly shuffle the data into an estimation sample of size $n_1 = 132$ and an independent evaluation sample of size $n_2 = 5$. Given the small evaluation sample size, we create 1,000 such random splits, compute the out-of-sample prediction for each split, and then summarise results over all 1,000 splits to alleviate concerns that results from a single split may not be representative. Summarising, the average predictive efficiency gain for the cross-validated kernel estimator was 8.4% relative to the parametric estimator, averaged over all 1,000 random shuffles. Of more direct interest is the ability of cross-validation to remove 'irrelevant variables' by assigning a bandwidth close to the permissible upper bound. Table 10 presents the median bandwidths over the 1,000 splits along with the 5th and 95th percentiles (upper bounds are given in square brackets).

Table 10. Median cross-validated bandwidth values over the 1,000 splits, with their 5th and 95th percentiles in parentheses. Numbers in square brackets represent maximum bandwidths, and so are the respective values of $(r_j - 1)/r_j$.

$\hat{\lambda}_1^{\text{med}}$ [0.50]	$\hat{\lambda}_2^{\text{med}}$ [0.75]	$\hat{\lambda}_3^{\text{med}}$ [0.92]	$\hat{\lambda}_4^{\text{med}}$ [0.96]	$\hat{\lambda}_5^{\text{med}}$ [0.98]	$\hat{\lambda}_6^{\text{med}}$ [0.50]
0.50	0.28	0.01	0.87	0.96	0.50
(0.50, 0.50)	(0.18, 0.36)	(0.00, 0.14)	(0.76, 0.92)	(0.72, 0.97)	(0.20, 0.50)

It can be seen from Table 10 that variables 1, 5, and 6 are effectively removed from the nonparametric estimator, indicating that cell type and Karnofsky score (variables 2 and 3) are deemed the most "relevant" by the cross-validation criterion.[2]

It would appear that, in small-sample settings involving a large number of covariates, the proposed estimator performs well for this popular dataset in terms of its predictive performance on independent evaluation data, particularly when compared with a common parametric specification.

6. PROOF OF THEOREM 4.1

We shall assume there are one or more irrelevant components, i.e. $p_1 < p$ or $q_1 < q$ or both, since otherwise the argument is much simpler. In this context, put

$$\nu_k(x) = E\left(\left[\left\{ \prod_{j=p_1+1}^{p} \frac{1}{h_j} K\left(\frac{x_j^{\mathrm{c}} - X_{1j}^{\mathrm{c}}}{h_j} \right) \right\} \prod_{j=q_1+1}^{q} \left\{ (1-\lambda_j)\, \rho_j^{N_{1j}(x)} \right\} \right]^k \right). \quad (6.1)$$

Step (i): Taylor expansion. Let $\mu_{f,-} = E(\hat{f}_{-i})$, $\mu_{G,-} = E(\widehat{G}_{-i})$, $\mu_{m,-} = E(\widehat{m}_{-i})$, $\Delta_{m,-i} = \widehat{m}_{-i} - \mu_{m,-}$, $\Delta_{f,-i} = \hat{f}_{-i} - \mu_{f,-}$ and $\Delta_{G,-i} = \widehat{G}_{-i} - \mu_{G,-}$. In this step we shall show that the following first-order Taylor expansion (i.e. expansion to linear terms) is valid:

$$\frac{\widehat{G}_{-i}(x)\, w(x^{\mathrm{c}})}{\widehat{m}_{-i}(x)^2} = \frac{\mu_{G,-}(x)\, w(x^{\mathrm{c}})}{\mu_{m,-}(x)^2} \left\{ 1 - \frac{2\,\Delta_{m,-i}(x)}{\mu_{m,-}(x)} \right\} + \frac{\Delta_{G,-i}(x)\, w(x^{\mathrm{c}})}{\mu_{m,-}(x)^2} + o_p\big\{(nH)^{-1}\big\}, \quad (6.2)$$

$$\frac{\hat{f}_{-i}(x,y)\, w(x^{\mathrm{c}})}{\widehat{m}_{-i}(x)} = \frac{\mu_{f,-}(x,y)\, w(x^{\mathrm{c}})}{\mu_{m,-}(x)} \left\{ 1 - \frac{\Delta_{m,-i}(x)}{\mu_{m,-}(x)} \right\} + \frac{\Delta_{f,-i}(x,y)\, w(x^{\mathrm{c}})}{\mu_{m,-}(x)} + o_p\big\{(nH)^{-1}\big\}, \quad (6.3)$$

uniformly in

$$\begin{aligned} &1 \le i \le n \text{ and } (x,y) \text{ such that } x = (x^{\mathrm{c}}, x^{\mathrm{d}}), \text{ with } x^{\mathrm{d}} \text{ in the support} \\ &\text{of the distribution of } X^{\mathrm{d}};\ h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q \text{ such that } h \le n^{-\epsilon} \\ &\text{and } n^{\epsilon-1} \le H \le n^{-\epsilon};\ \text{and uniformly in } \lambda_j \in \Lambda_j \text{ for } 1 \le j \le q; \end{aligned} \quad (6.4)$$

for arbitrary but fixed $\epsilon \in (0, \frac{1}{2})$. Initially we shall work with the regular, rather than the leave-one-out, estimators.

[2] The Karnofsky score measures patient performance of activities of daily living. The score has proven useful not only for following the course of the illness (usually progressive deficit and ultimately death), but also a prognosticator: patients with the highest (best) Karnofsky scores at the time of tumour diagnosis have the best survival and quality of life over the course of their illness.

Indeed, defining $\mu_m = E(\widehat{m}) = \mu_{m,-}$, $\mu_f = E(\hat{f}) = \mu_{f,-}$, $\mu_G = E(\widehat{G})$, $\Delta_m = \widehat{m} - \mu_m$, $\Delta_f = \hat{f} - \mu_f$ and $\Delta_G = \widehat{G} - \mu_G$, we shall show that for some $\delta > 0$ and all $C > 0$,

$$P\biggl(\sup \biggl| \frac{\widehat{G}(x)\,w(x^{\rm c})}{\widehat{m}(x)^2} - \biggl[\frac{\mu_G(x)\,w(x^{\rm c})}{\mu_m(x)^2} \biggl\{ 1 - \frac{2\,\Delta_m(x)}{\mu_m(x)} \biggr\} + \frac{\Delta_G(x)\,w(x^{\rm c})}{\mu_m(x)^2} \biggr] \biggr| > n^{-\delta}\,(nH)^{-1} \biggr) = O\bigl(n^{-C}\bigr)\,, \tag{6.5}$$

$$P\biggl(\sup \biggl| \frac{\hat{f}(x,y)\,w(x^{\rm c})}{\widehat{m}(x)} - \biggl[\frac{\mu_f(x,y)\,w(x^{\rm c})}{\mu_m(x)} \biggl\{ 1 - \frac{\Delta_m(x)}{\mu_m(x)} \biggr\} + \frac{\Delta_f(x,y)\,w(x^{\rm c})}{\mu_m(x)} \biggr] \biggr| > n^{-\delta}\,(nH)^{-1} \biggr) = O\bigl(n^{-C}\bigr)\,, \tag{6.6}$$

where the supremum is over all $x, y, h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ prescribed by (6.4). Since C is arbitrarily large then we may also take the supremum over any polynomially large number of copies of the sample $\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, regardless of whether the copies are independent of one another. (Of course, each copy sample must have the same distribution as $\mathcal{Z}$, and in particular must be itself a random sample.) Now, there is only a polynomially large number of values of i, and each leave-one-out sample $\mathcal{Z} \backslash \{(X_i, Y_i)\}$ is a random $(n-1)$-sample, and so (6.5) and (6.6) imply (6.2) and (6.3) for a sample of size $n+1$ rather than n.

Therefore it suffices to derive (6.5) and (6.6). For this it is sufficient, by Taylor expansion, to show that for some $\delta > 0$ and all $C > 0$,

$$P\biggl[\sup \biggl| \frac{\Delta_m(x)\,w(x^{\rm c})}{\nu_1(x)} \biggr| \biggl\{ \biggl| \frac{\Delta_f(x,y)}{\nu_1(x)} \biggr| + \biggl| \frac{\Delta_G(x)}{\nu_1(x)^2} \biggr| \biggr\} > n^{-\delta}\,(nH)^{-1} \biggr] = O\bigl(n^{-C}\bigr)\,, \tag{6.7}$$

where again the supremum is over $x, y, h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ determined by (6.4).

Let $\widehat{G} = \widehat{G}_{\rm diag} + \widehat{G}_{\rm offd}$ denote the decomposition of $\widehat{G}$ into its diagonal and off-diagonal components, and observe that

$$\widehat{m}(x) = \frac{1}{n} \sum_{i=1}^n U_i(x)\,, \qquad \hat{f}(x,y) = \frac{1}{nh} \sum_{i=1}^n U_i(x)\,L\biggl(\frac{y - Y_i}{h} \biggr)\,,$$

$$\widehat{G}_{\rm diag}(x) = \frac{\kappa_L}{n^2 h} \sum_{i=1}^n U_i(x)^2\,, \quad \widehat{G}_{\rm offd}(x) = \frac{1}{n^2 h} \sum_{i_1 \neq i_2} \sum U_{i_1}(x)\,U_{i_2}(x)\,L_2\biggl(\frac{Y_{i_1} - Y_{i_2}}{h} \biggr)\,,$$

where $\kappa_L = \int L^2$, L_2 denotes the convolution of L with itself,

$$U_i(x) = \biggl\{ \prod_{j=1}^p \frac{1}{h_j}\,K\biggl(\frac{x_j^{\rm c} - X_{ij}^{\rm c}}{h_j} \biggr) \biggr\} \prod_{j=1}^q \Bigl\{ (1 - \lambda_j)\,\rho_j^{N_{ij}(x)} \Bigr\}\,,$$

and $\rho_j = \lambda_j/\{(1-\lambda_j)\,(r_j-1)\}$. Put $\Delta_{G\,\mathrm{diag}} = \widehat{G}_{\mathrm{diag}} - E\widehat{G}_{\mathrm{diag}}$, $\Delta_{G\,\mathrm{offd}} = \widehat{G}_{\mathrm{offd}} - E\widehat{G}_{\mathrm{offd}}$ and $H_k = \prod_j \max(h_j^{k-1}, h_j^k)$, for integers $k \geq 1$. Uniformly in x, y, h, $h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$, we have:

$$H_{2k}\, E\big\{U_1(x)^{2k}\big\} + H_{2k}\, \max\big(h^{2k-1}, h^{2k}\big)\, E\bigg\{U_1(x)\,\frac{1}{h}\,L\bigg(\frac{y-Y_1}{h}\bigg)\bigg\}^{2k} = O(1)\,.$$

Therefore, by Rosenthal's inequality (Hall and Heyde, 1980, p. 23),

$$\begin{aligned}
E\big\{\Delta_m(x)^{2k}\big\} &\leq C_k\,\big\{(nH_2)^{-k} + n^{1-2k}\,H_{2k}^{-1}\big\}\,,\\
E\big\{\Delta_f(x,y)^{2k}\big\} &\leq C_k\,\Big[\big\{nH_2\,\max\big(h,h^2\big)\big\}^{-k} + n^{1-2k}\,H_{2k}^{-1}\big\{\,\max\big(h^{2k-1},h^{2k}\big)\big\}^{-1}\Big],\\
E\big\{\Delta_{G\,\mathrm{diag}}(x)^{2k}\big\} &\leq C_k\,\big\{\big(n^3H_4\big)^{-k} + n^{1-4k}\,H_{4k}^{-1}\big\}\,h^{-2k}\,,
\end{aligned}$$

uniformly in the same sense as before, where, here and below, C_k denotes a generic positive constant depending only on k (as well as on the kernels K and L and the distribution of (X,Y)).

Since the conditional density of X^{c}, given $X^{\mathrm{d}} = x^{\mathrm{d}}$, is bounded away from zero on $\operatorname{supp} w$, for each x^{d} in the support of the distribution of X^{d}, then

$$\nu_k(x) \prod_{j=p_1+1}^{p} \max\big(h_j^{k-1}, h_j^k\big)$$

is bounded away from zero and infinity, uniformly in $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$ with $x^{\mathrm{c}} \in \operatorname{supp} w$ and x^{d} in the support of the distribution of X^{d}. Therefore, if we define

$$J = \bigg(\prod_{j=1}^{p_1} h_j\bigg) \prod_{j=p_1+1}^{p_2} \min(h_j, 1)\,,$$

then $\nu_1^{2k} H_{2k}$ is bounded below by a constant multiple of J^{2k-1}. Hence, for another constant C_k,

$$E\bigg|\frac{\Delta_m(x)}{\nu_1(x)}\bigg|^{2k} \leq C_k\,\big\{(nJ)^{-k} + (nJ)^{1-2k}\big\}\,, \tag{6.8}$$

$$E\bigg|\frac{\Delta_f(x,y)}{\nu_1(x)}\bigg|^{2k} \leq C_k\,\big\{\big(nhJ\big)^{-k} + (nhJ)^{1-2k}\big\}\,, \tag{6.9}$$

$$\begin{aligned}
E\bigg|\frac{\Delta_{G\,\mathrm{diag}}(x)}{\nu_1(x)^2}\bigg|^{2k} &\leq C_k\,\big\{(nJ)^{-3k} + \big(nJ\big)^{1-4k}\big\}\,h^{-2k}\\
&\leq C_k\,\big\{(nhJ)^{-3k} + \big(nhJ\big)^{1-4k}\big\}\,,
\end{aligned} \tag{6.10}$$

the latter inequality following since $h \leq 1$.

Note that $H = hJ$. We shall show in step (ii) below that for any $\delta > 0$ we have for all sufficiently large k,

$$E\left|\frac{\Delta_{G\,\mathrm{offd}}(x)}{\nu_1(x)^2}\right|^{2k} \le C_k(\delta)\, n^{k\delta}\left\{(nH)^{-k} + (nH)^{1-4k}\right\}. \tag{6.11}$$

An heuristic argument leading to (6.11) is the following:

$$\begin{aligned} E\left|\frac{\Delta_{G\,\mathrm{offd}}(x)}{\nu_1(x)^2}\right|^{2k} &\le \text{const.}\left[\mathrm{var}\left\{\frac{\Delta_{G\,\mathrm{offd}}(x)}{\nu_1(x)^2}\right\}\right]^{k-c_1} \\ &\le \text{const.}\,(nH)^{c_2-2k} \le \text{const.}\,(nH)^{-k}, \end{aligned} \tag{6.12}$$

for constants $c_1, c_2 > 0$. Result (6.12) can be verified by lengthy but direct calculation, treating all the permutations that make a nonvanishing contribution to the expansion of the $2k$th moment on the left-hand side. We shall give instead an argument, based on Rosenthal's inequality, sufficient to give the bound at (6.11).

Combining (6.8)–(6.11), noting that $h = O(n^{-\epsilon})$ for some $\epsilon > 0$, and applying Markov's inequality, we deduce that

$$P\{D(\xi) > n^{\delta}\,(nH)^{-1/2}\} = O\big(n^{-C}\big) \tag{6.13}$$

for all $\delta, C > 0$, where $\xi = (x, y, h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ and

$$D(\xi) = \left|\frac{\Delta_m(x)}{h^{1/2}\,\nu_1(x)}\right| + \left|\frac{\Delta_f(x,y)}{\nu_1(x)}\right| + \left|\frac{\Delta_{G\,\mathrm{diag}}(x)}{\nu_1(x)^2}\right| + \left|\frac{\Delta_{G\,\mathrm{offd}}(x)}{\nu_1(x)^2}\right|.$$

The remainder in (6.13) is of the stated order uniformly in ξ prescribed by (6.4). Since C in (6.13) is arbitrarily large then the same result holds uniformly in any set, of size no larger than a polynomial in n, of values of ξ prescribed by (6.4). That is, if $\mathcal{S}$ is any such set of values of ξ then

$$P\left\{\sup_{\xi\in\mathcal{S}} D(\xi) > n^{\delta}\,(nH)^{-1/2}\right\} = O\big(n^{-C}\big) \tag{6.14}$$

for all $\delta, C > 0$. The regularity conditions asserted in Theorem 4.1 imply that each of $h, h_1, \ldots, h_p$ is no smaller than n^{-C_1}, and no larger than n^{C_2}, for constants $C_1, C_2 > 0$. Furthermore, the functions K and L are compactly supported and Hölder continuous. Therefore, taking a polynomially fine mesh of vectors ξ prescribed by (6.4) we deduce that (6.14) continues to hold if $\mathcal{S}$ is replaced by the class of all ξ satisfying (6.4).

Result (6.7) follows from (6.14), on noting the factor $h^{1/2}$ in the denominator of the first term on the right-hand side in the definition of $D(\xi)$, and recalling that, in (6.4), we imposed the constraint $h \le n^{-\epsilon}$ for some $\epsilon > 0$.

Step (ii): Derivation of (6.11). Let $\mathcal{F}_i$ denote the σ-field generated (X_i, Y_i), and for $i_1 < i_2$ put

$$Q_{i_1 i_2}(x) = U_{i_1}(x)\, U_{i_2}(x)\, L_2\left(\frac{Y_{i_1} - Y_{i_2}}{h}\right), \quad R_{i_1} = E(Q_{i_1 i_2} \,|\, \mathcal{F}_{i_1}),$$

$$S_{i_1 i_2} = Q_{i_1 i_2} - R_{i_1} - R_{i_2} + E(R_1), \quad S = \sum_{i_1 < i_2} \sum S_{i_1 i_2},$$

$$\Delta^{(1)}_{G\,\mathrm{offd}} = 2\left(n^2 h\right)^{-1} S, \quad \Delta^{(2)}_{G\,\mathrm{offd}} = \frac{2\,(n-1)}{n^2 h} \sum_{i=1}^{n} (R_i - E R_1).$$

In this notation,

$$\Delta_{G\,\mathrm{offd}} = \Delta^{(1)}_{G\,\mathrm{offd}} + \Delta^{(2)}_{G\,\mathrm{offd}}.$$

The variables R_i are independent and identically distributed, have zero mean, and satisfy $E(R_i^{2k}) \le \text{const.}\, h^{2k} H_{2k}^{-1}$. Therefore, by Rosenthal's inequality,

$$E\left[\left\{\Delta^{(2)}_{G\,\mathrm{offd}}(x)\right\}^{2k}\right] \le \text{const.}\left\{(nH_2)^{-k} + n^{1-2k}\, H_{2k}^{-1}\right\}.$$

Hence, since $\nu_1(x)^{2k} H_{2k}$ is bounded below by a constant multiple of J^{2k-1}, uniformly in $x \in \operatorname{supp} w$, then

$$E\left|\frac{\Delta^{(2)}_{G\,\mathrm{offd}}(x)}{\nu_1(x)^2}\right|^{2k} \le \text{const.}\left\{(nJ)^{-k} + \left(nJ\right)^{1-2k}\right\}.$$

Therefore it suffices to prove the version of (6.11) for $\Delta^{(1)}_{G\,\mathrm{offd}}$, for which it is adequate to show that for each $\delta > 0$, and all sufficiently large k,

$$E\left|\Delta^{(1)}_{G\,\mathrm{offd}}(x)\right|^{2k} \le C_k(\delta)\, n^{k\delta}\left\{(nH_2 h)^{-k} + (nh)^{-4k}\, H_{4k}^{-1}\right\}. \tag{6.15}$$

We shall give next the first step in the argument leading to (6.15), and then outline the subsequent steps.

Let $\mathcal{G}_i$ denote the σ-field generated by (X_{i_1}, Y_{i_1}) for $i_1 \le i$. Now, $T \equiv S$ is a degenerate U-statistic, expressible as a martingale with zero mean and martingale differences $T_{i_2} = \sum_{1 \le i_1 \le i_2 - 1} S_{i_1 i_2}$:

$$T = \sum_{i_1 < i_2} \sum S_{i_1 i_2} = \sum_{i=2}^{n} T_i, \tag{6.16}$$

where $E(T_i|\mathcal{G}_{i-1}) = 0$. By Rosenthal's inequality,

$$E\big(T^{2k}\big) \le C_k \left[E\Big\{ \sum_{i=2}^{n} E\big(T_i^2 \big| \mathcal{G}_{i-1}\big) \Big\}^k + \sum_{i=2}^{n} E\big(T_i^{2k}\big) \right]. \tag{6.17}$$

In order to prove (6.15) it suffices, in view of (6.17), to show that for each $\delta > 0$, and all sufficiently large k,

$$E\Big\{ \sum_{i=2}^{n} E\big(T_i^2 \big| \mathcal{G}_{i-1}\big) \Big\}^k + \sum_{i=2}^{n} E\big(T_i^{2k}\big) \le C_k(\delta)\, n^{k\delta} \left\{ \big(n^3 h/H_2\big)^k + H_{4k}^{-1} \right\}. \tag{6.18}$$

In deriving (6.18) it is relatively straightforward to bound the second term on the left-hand side. Indeed, we shall prove shortly that

$$\max_{2 \le i \le n} E\big(T_i^{2k}\big) \le \text{const.} \left\{ \big(n^3 h/H_2\big)^k + nh\, H_{4k}^{-1} \right\}, \tag{6.19}$$

whence we obtain,

$$\sum_{i=2}^{n} E\big(T_i^{2k}\big) \le \text{const.}\, n \left\{ \big(n^3 h/H_2\big)^k + nh\, H_{4k}^{-1} \right\}. \tag{6.20}$$

This establishes the bound at (6.18) for the second term on the left there. A crude argument now shows that the first term on the left-hand side of (6.18) admits the same inequality, but with the factor n replaced by n^k: for each $\delta > 0$, and all sufficiently large k,

$$\begin{aligned} E\Big\{ \sum_{i=2}^{n} E\big(T_i^2 \big| \mathcal{G}_{i-1}\big) \Big\}^k &\le n^k \max_{2 \le i \le n} E\big(T_i^{2k}\big) \\ &\le \text{const.}\, n^k \left\{ \big(n^3 h/H_2\big)^k + nh\, H_{4k}^{-1} \right\} \le \text{const.}\, n^{k\delta + k}\, b_k\,, \end{aligned} \tag{6.21}$$

where $b_k = (n^3 h/H_2)^{-k} + H_{4k}^{-1}$.

However, if we express $E(T_i^2|\mathcal{G}_{i-1})$ as a martingale in the same way we expressed T, and iterate the argument leading to (6.19)–(6.21), we may reduce the factor $n^{k\delta+k}$ in (6.21) first to $n^{k\delta+(k/2)}$, then to $n^{k\delta+(k/4)}$, and so on. Two paragraphs below we shall give this argument.

First we derive (6.19), however. Conditional on $\mathcal{F}_{i_2}$, T_{i_2} is a sum of independent random variables with zero means, and so by Rosenthal's inequality again,

$$E\big(T_{i_2}^{2k}\big) \le C_k \left[E\Big\{ \sum_{i_1=1}^{i_2-1} E\big(S_{i_1 i_2}^2 \big| \mathcal{F}_{i_2}\big) \Big\}^k + \sum_{i_1=1}^{i_2-1} E\big(S_{i_1 i_2}^{2k}\big) \right]. \tag{6.22}$$

Now, $E(S_{i_1 i_2}^{2k}) \le \text{const.}\, H_{2k}^{-2} h$, and $E(S_{i_1 i_2}^2 | \mathcal{F}_{i_2}) = Z_{i_2} H_2^{-1} h$, where $E(Z_{i_2}^k) \le \text{const.}\, H_{2k}^{-1}$, uniformly in $2 \le i_2 \le n$. Therefore,

$$E\Big\{ \sum_{i_1=1}^{i_2-1} E\big(S_{i_1 i_2}^2 | \mathcal{F}_{i_2}\big) \Big\}^k \le \text{const.}\, (nh/H_2)^k \, H_{2k}^{-1} \,,$$

uniformly in $2 \le i_2 \le n$. Hence, by (6.22),

$$\max_{2 \le i \le n} E\big(T_i^{2k}\big) \le \text{const.}\, \big\{ (nh/H_2)^k \, H_{2k}^{-1} + nh \, H_{2k}^{-2} \big\} \,. \tag{6.23}$$

Since $H_{2k} \ge J^{2k-1}$, $H_{2k}^2 \ge H_{4k}$, $J \ge n^{-1}$ and

$$(nh/H_2)^k \, J^{1-2k} = \big(n^3 h/H_2\big)^k \, (nJ)^{1-2k} \, n^{-1} \le \big(n^3 h/H_2\big)^k \,,$$

then (6.19) follows from (6.23).

Returning to (6.21), we describe the method for sharpening that bound. The technique is iterative, and at the ℓth step involves bounding moments of order $t(\ell) = k/2^\ell$, rather than of order $2k$ or k. Reflecting this, let us take $k = 2^r$ for an integer $r \ge 0$, and let $0 \le \ell \le r-1$. We shall derive the version of (6.21) in which $n^{k\delta+k} \, b_k$, on the right-hand side, is replaced by $n^{k\delta} \, b_k$. This result, together with (6.20), is sufficient to give (6.18) and so to complete the proof of (6.15) and hence also the proof of (6.11).

Suppose that for $\ell \ge 0$ we have a statistic $T^{(\ell)}$ that can be written as

$$T^{(\ell)} = \sum_{i_1 < i_2} \sum S_{i_1 i_2}^{(\ell)} = \sum_{i=2}^n T_i^{(\ell)} \,, \tag{6.24}$$

say, where $T_{i_2}^{(\ell)} = \sum_{1 \le i_1 \le i_2 - 1} S_{i_1 i_2}^{(\ell)}$ and, for each ℓ, $S_{i_1 i_2}^{(\ell)}$ is a symmetric, deterministic function of (X_{i_1}, Y_{i_1}) and (X_{i_2}, Y_{i_2}), satisfying $E(S_{i_1 i_2}^{(\ell)} | \mathcal{F}_i) = 0$ for $i = i_1, i_2$. Therefore, (6.24) expresses $T^{(\ell)}$ as a degenerate U-statistic and as a martingale, with differences $T_i^{(\ell)}$. We start with $T^{(0)} = T = S$, given at (6.16). By Rosenthal's inequality,

$$E\big\{ (T^{(\ell)})^{2t(\ell)} \big\} \le C_k \left\{ E\left(\left[\sum_{i_2=2}^n E\left\{ \left(\sum_{i_1=1}^{i_2-1} S_{i_1 i_2}^{(\ell)} \right)^2 \,\middle|\, \mathcal{G}_{i_2-1} \right\} \right]^{t(\ell)} \right) + s_1^{(\ell)} \right\} , \tag{6.25}$$

where

$$s_1^{(\ell)} = \sum_{i_2=2}^n E\left(\sum_{i_1=1}^{i_2-1} S_{i_1 i_2}^{(\ell)} \right)^{2t(\ell)} .$$

The right-hand side here, being a sum of moments of sums of random variables which are independent conditional on (X_{i_2}, Y_{i_2}), is straightforward to accurately bound. Indeed, the argument leading to (6.20) gives, for $j = 1$ and $0 \le \ell \le r-1$,

$$s_j^{(\ell)} \le \text{const.}\, n^{k\delta}\, b_k\,. \tag{6.26}$$

The expectation on the right-hand side of (6.25) we deal with by noting that

$$\sum_{i_2=2}^{n} E\biggl\{ \Bigl(\sum_{i_1=1}^{i_2-1} S_{i_1 i_2}^{(\ell)} \Bigr)^2 \,\Big|\, \mathcal{G}_{i_2-1} \biggr\} = T^{(\ell+1)} + \sum_{i_1=1}^{n-1} (n-i_1)\, E\bigl\{ (S_{i_1,i_1+1}^{(\ell)})^2 \,\big|\, \mathcal{G}_{i_1} \bigr\} + 2\,(n-1) \sum_{i=1}^{n} \bigl(R_i^{(\ell+1)} - E R_1^{(\ell+1)} \bigr)\,, \tag{6.27}$$

where $T^{(\ell+1)} = \sum \sum_{i_1<i_2} S_{i_1 i_2}^{(\ell+1)}$ and, for $i_1 < i_2$,

$$S_{i_1 i_2}^{(\ell+1)} = Q_{i_1 i_2}^{(\ell+1)} - R_{i_1}^{(\ell+1)} - R_{i_2}^{(\ell+1)} + E\bigl(R_1^{(\ell+1)}\bigr)\,,$$
$$Q_{i_1 i_2}^{(\ell+1)} = 2\,(n-i_2+1)\, E\Bigl(S_{i_1,i_2+1}^{(\ell)}\, S_{i_2,i_2+1}^{(\ell)} \,\Big|\, \mathcal{G}_{i_2} \Bigr)$$

and $R_{i_1}^{(\ell+1)} = E(Q_{i_1 i_2}^{(\ell+1)} | \mathcal{F}_{i_1})$. Let $s_2^{(\ell)}$ denote the $t(\ell)$th moment of the sum of the two series on the right-hand side of (6.27). The series are of independent random variables, and so again (6.26) is relatively straightforward to derive. The quantity $T^{(\ell+1)}$ is analogous to $T^{(\ell)}$, and can be expressed as a degenerate U-statistic and as a martingale, just as was $T^{(\ell)}$ at (6.24). Note too that

$$E\bigl\{ (T^{(\ell)})^{2t(\ell)} \bigr\} \le \text{const.} \Bigl[E\bigl\{ (T^{(\ell+1)})^{t(\ell)} \bigr\} + s_1^{(\ell)} + s_2^{(\ell)} \Bigr]\,. \tag{6.28}$$

Importantly, in passing from the left- to the right-hand side of (6.28) the exponent $2t(\ell)$ is halved. We iterate (6.28) until $\ell = r-1$, i.e. $t(\ell) = 2$. The argument leading to (6.21) gives $E\{(T^{(\ell)})^{2t(\ell)}\} \le \text{const.}\, n^{k\delta + t(\ell)}\, b_k$, for each $0 \le \ell \le r-1$ and δ. Taking $\ell = r-1$, in which case $t(\ell) = 2$ and the moment on the right-hand side of (6.28) is just the mean square, we show that $E\{(T^{(r)})^2\} \le \text{const.}\, n^{k\delta}\, b_k$. Combining this result with (6.26) and (6.28), for general ℓ, we deduce that $E\{(T^{(0)})^{2k}\} \le \text{const.}\, n^{k\delta}\, b_k$, which is the desired form of (6.21) in which the factor $n^{k\delta+k}$ there is replaced by $n^{k\delta}$.

Step (iii): Separation into stochastic and non-stochastic terms. Noting (6.2), and the fact that $\mu_f = \mu_{f,-i}$, $\mu_m = \mu_{m,-i}$ and $\mu_G = \mu_{G,-i} + O(n^{-1})$, we see that

$$\text{CV} = I_1 - 2\, I_2 + \Delta_1 + \Delta_2 + o_p\bigl\{ (nH)^{-1} \bigr\}\,, \tag{6.29}$$

uniformly in values of the smoothing parameters satisfying (6.4), where $\Delta_1 = S_1 - E(S_1)$,

$$I_1 = \int \frac{\mu_G(x)}{\mu_m(x)^2}\, m\, dW(x)\,, \quad I_2 = \int \frac{\mu_f(x,y)}{\mu_m(x)}\, f(x,y)\, dW(x)\,,$$
$$S_1 = \frac{1}{n}\sum_{i=1}^{n} w(X_i^{\mathrm{c}}) \left\{ \frac{\mu_G(X_i)}{\mu_m(X_i)^2} - 2\,\frac{\mu_f(X_i,Y_i)}{\mu_m(X_i)} \right\},$$
$$\Delta_2 = \frac{1}{n}\sum_{i=1}^{n} w(X_i^{\mathrm{c}}) \left\{ \frac{\Delta_{G,-i}(X_i)}{\mu_m(X_i)^2} - 2\,\frac{\Delta_{f,-i}(X_i,Y_i)}{\mu_m(X_i)} - 2\,\frac{\mu_G(X_i)\,\Delta_{m,-i}(X_i)}{\mu_m(X_i)^3} + 2\,\frac{\mu_f(X_i,Y_i)\,\Delta_{m,-i}(X_i)}{\mu_m(X_i)^2} \right\}. \tag{6.30}$$

Note that $I_1 - 2\,I_2 = E(S_1)$, and define

$$D = \int \{\mu_g(y|x) - g(y|x)\}^2\, m(x)\, dW(x)\, dy\,. \tag{6.31}$$

We shall refer to $I_1 - 2\,I_2$ as the non-stochastic term in an expansion of CV, and call Δ_1 and Δ_2 the stochastic terms. We shall bound the latter in step (v), showing that for random variables Q_1 and Q_2 not depending on the smoothing parameters, for some $\delta > 0$ and all $C > 0$, and for $j = 1, 2$,

$$P\Big[\sup \Big| \Delta_j(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) - Q_j \Big| > n^{-\delta} \left\{ (nH)^{-1} + D \right\} \Big] = O\big(n^{-C}\big)\,, \tag{6.32}$$

where the supremum is over all values of $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ prescribed by (6.4).

Step (iv): Expansion of $I_1 - 2\,I_2$, and proof of theorem. Recall from step (i) that $\mu_f(x,y) = E\{K(x,X)\,L(y,Y)\}$ and $\mu_m(x) = E\{K(x,X)\}$, and put $\mu_{f2}(x,y) = E\{K(x,X)^2\,L(y,Y)^2\}$, $\mu_g = \mu_f/\mu_m$ and

$$I_3 = \int g(y|x)^2\, m(x)\, dW(x)\, dy\,,$$

the latter not depending on any smoothing parameters. Note that

$$I_1 = \big(1 - n^{-1}\big) \int \mu_g(y|x)^2\, m(x)\, dW(x)\, dy + n^{-1} \int \frac{\mu_{f2}(x,y)}{\mu_m(x)^2}\, m(x)\, dW(x)\, dy\,.$$

It can be shown that the first integral on the right-hand side is bounded uniformly in $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4). Therefore,

$$I_1 - 2\,I_2 + I_3 = D + n^{-1} \int \frac{\mu_{f2}(x,y)}{\mu_m(x)^2}\, m(x)\, dW(x)\, dy + O\big(n^{-1}\big)\,, \tag{6.33}$$

uniformly in $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4).

Define $\tau = \nu_2/\nu_1^2$, where ν_1 and ν_2 are as at (6.1). Recall that "bar" notation indicates that irrelevant components have been dropped from functions or vectors. In particular, let $\bar{\mu}_f$, $\bar{\mu}_{f2}$ $\bar{\mu}_g$ and $\bar{\mu}_m$ denote the versions of μ_f, μ_{f2}, μ_g and μ_m, respectively, obtained by eliminating the last $p_2 = p - p_1$ components of x^{c}, and the last $q_2 = q - q_1$ components of x^{d}. Using the fact that these components are totally independent of the others it may be shown that $\mu_f = \bar{\mu}_f\, \nu_1$, $\mu_{f2} = \bar{\mu}_{f2}\, \nu_2$ and $\bar{\mu}_m = \mu_m\, \nu_1$. Therefore, $\mu_g = \mu_f/\mu_m = \bar{\mu}_g$ and $\mu_{f2}(x, y)/\mu_m(x)^2 = \tau(x)\, \bar{\mu}_{f2}(\bar{x}, y)/\bar{\mu}_m(\bar{x})^2$. Note too that $g(y|x) = \bar{g}(y|\bar{x})$. Substituting these results into (6.33) we deduce that

$$I_1 - 2\, I_2 + I_3 = D + n^{-1} \int \frac{\bar{\mu}_{f2}(\bar{x}, y)}{\bar{\mu}_m(\bar{x})^2}\, \tau(x)\, m(x)\, dW(x)\, dy + O\big(n^{-1}\big)\,, \qquad (6.34)$$

uniformly in $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4).

If the values of $h, h_1, \ldots, h_{p_1}, \lambda_1, \ldots, \lambda_{q_1}$ that, along with the other smoothing parameters, minimise CV, do not all converge in probability to zero, then, by (4.6), D (defined at (6.31)) does not converge to zero. This result; a subsequence argument; the fact that D is bounded uniformly in the smoothing parameters; the property that the "stochastic terms" isolated in step (iii) all converge to zero uniformly in the smoothing parameters, up to terms that do not depend on those parameters (see (6.32)); formulae (6.29) and (6.34); and the fact that I_3 does not depend on the smoothing parameters; together imply that for some $\delta > 0$, the probability that the minimum of CV $+ I_3$, over the smoothing parameters, exceeds δ, does not converge to zero as $n \to \infty$.

However, choosing $h, h_1, \ldots, h_{p_1}$ to be of size $n^{-1/(p_1+5)}$, and $\lambda_1, \ldots, \lambda_{q_1}$ to be of size $n^{-2/(p_1+5)}$; taking $h_{p_1+1}, \ldots, h_p$ to diverge to infinity, and λ_j to converge to $(r_j - 1)/r_j$ for $q_1 + 1 \le j \le q$; and using the fact that the "stochastic terms" all converge to zero up to quantities that do not depend on the smoothing parameters ((6.32) again); it may similarly be shown that CV $+ I$ converges in probability to zero. This contradicts the result obtained in the previous paragraph, and thus demonstrates that:

> at the minimum of CV, the smoothing parameters $h, h_1, \ldots, h_{p_1}$, $\lambda_1, \ldots, \lambda_{q_1}$, for the relevant components of X, all converge in probability to zero. (6.35)

For any sequence $\eta_n \downarrow 0$,

$$\left| \frac{\int \frac{\bar{\mu}_{f2}(x,y)}{\bar{\mu}_m(x)^2}\, \tau(x)\, m(x)\, dW(x)\, dy}{\frac{\kappa^p\, \kappa_L}{h\, h_1 \ldots h_{p_1}} \int g(y|x)\, \tau(x)\, dW(x)\, dy} - 1 \right| \to 0\,, \tag{6.36}$$

uniformly in values $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4) and such that $h, h_1, \ldots, h_{p_1}$ lie in the interval $(0, \eta_n]$. The result derived in the previous paragraph therefore implies that (6.36) holds, with convergence in probability, if $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ are chosen to minimise CV.

Let $\bar{\chi}_1$ denote the version of $\bar{\chi}$, at (4.2), that is obtained if we remove the contribution to that quantity from the variance component of weighted mean integrated squared error:

$$\begin{aligned}\bar{\chi}_1(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) = \bar{\chi}(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\ - \frac{\kappa^{p_1}\, \kappa_L}{a_0 \ldots a_{p_1}} \sum_{\bar{x}^{\rm d}} \int \bar{f}(\bar{x}, y)\, \bar{w}(\bar{x}^{\rm c}, \bar{x}^{\rm d})\, d\bar{x}^{\rm c}\, dy\,.\end{aligned}$$

Put $a_j = n^{1/(p_1+1)} h_j$ and $b_j = n^{2/(p_1+1)} \lambda_j$. We may deduce, by Taylor expansion, that

$$\begin{aligned}&\int \{\bar{\mu}_g(y|\bar{x}) - g(y|\bar{x})\}^2\, m(x)\, dW(x)\, dy \\ &\quad = \int \{\bar{\mu}_f(\bar{x}, y) - \bar{\mu}_m(\bar{x})\, g(y|\bar{x})\}^2\, \frac{dW(x)\, dy}{m(x)} \\ &\qquad + o\Big(h^2 + h_1^2 + \ldots + h_{p_1}^2 + \lambda_1 + \ldots + \lambda_{q_1}\Big) \\ &\quad = n^{-2/(p+1)}\, \bar{\chi}_1(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\ &\qquad + o\Big(h^2 + h_1^2 + \ldots + h_{p_1}^2 + \lambda_1 + \ldots + \lambda_{q_1}\Big)\,,\end{aligned} \tag{6.37}$$

uniformly in $h, h_1, \ldots, h_{p_1}, \lambda_1, \ldots, \lambda_{q_1} \in (0, \eta_n]$, for any given sequence $\eta_n \downarrow 0$. Note that only h and the first p_1 h_j's, and the first q_1 λ_j's, feature in (6.37); none of the other smoothing parameters plays a role in any of the functions there.

Combining (6.34), (6.36) and (6.37) we deduce that

$$\begin{aligned}I_1 - 2\, I_2 + I_3 = n^{-2/(p+1)}\, \bar{\chi}(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\ + \{1 + o(1)\}\, \frac{\kappa\, \kappa_L}{n\, h\, h_1 \ldots h_{p_1}} \int \bar{g}(y|\bar{x})\, \{\tau(x) - 1\}\, dW(x)\, dy \\ + o\Big\{h^2 + h_1^2 + \ldots + h_{p_1}^2 + \lambda_1 + \ldots + \lambda_{q_1} + (nH)^{-1}\Big\}\,,\end{aligned} \tag{6.38}$$

uniformly in values of the smoothing parameters $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4) and such that $h, h_1, \ldots, h_{p_1}, \lambda_1, \ldots, \lambda_{q_1} \in (0, \eta_n]$. Results (6.29), (6.32) and

(6.38) imply that (6.38) continues to hold if we replace the left-hand side by CV, provided we add to the right-hand side a term Q that does not depend on the smoothing parameters:

$$\begin{aligned}\mathrm{CV} + I_3 = {} & n^{-2/(p+1)}\, \bar\chi(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\ & + \{1 + o_p(1)\}\, \frac{\kappa\, \kappa_L}{n\, h\, h_1 \ldots h_{p_1}} \int \bar g(y|\bar x)\, \{\tau(x) - 1\}\, dW(x)\, dy \\ & + Q + o_p\Big\{h^2 + h_1^2 + \ldots + h_{p_1}^2 + \lambda_1 + \ldots + \lambda_{q_1} + (nH)^{-1}\Big\}\,, \quad (6.39)\end{aligned}$$

uniformly in $h, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4) and such that $h, h_1, \ldots, h_{p_1}, \lambda_1, \ldots, \lambda_{q_1} \in (0, \eta_n]$.

Let $C > 0$ and $0 < \delta < \min(r_j - 1)/r_j$. Using properties (4.4) and (4.5) it may be proved that

$$H \int \bar g(y|\bar x)\, \{\tau(x) - 1\}\, dW(x)\, dy$$

is bounded away from zero and infinity uniformly in $h_{p_1+1}, \ldots, h_p$ and $\lambda_{q_1+1}, \ldots, \lambda_q$ such that $\inf(h_{p_1+1}, \ldots, h_p) \le C$ and

$$\max\big\{(r_j - 1)\, r_j^{-1} - \lambda_j \,:\, q_1 + 1 \le j \le q\big\} > \delta\,.$$

Therefore, (6.39) may be equivalently written as:

$$\begin{aligned}\mathrm{CV} + I_3 = {} & n^{-2/(p+1)}\, \bar\chi(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\ & + (nH)^{-1}\, P\big(h_{p_1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q\big) \\ & + Q + o_p\Big\{h^2 + h_1^2 + \ldots + h_{p_1}^2 + \lambda_1 + \ldots + \lambda_{q_1} + (nH)^{-1}\Big\}\,, \quad (6.40)\end{aligned}$$

where the deterministic function P is nonnegative and equals zero only in the limit as each of $h_{p_1}, \ldots, h_p$ diverges to infinity and each of $\lambda_{q_1+1}, \ldots, \lambda_q$ converges to the respective value of $(r_j - 1)/r_j$. Theorem 4.1 follows from (6.35) and (6.40), noting the uniqueness of the minimum of $\bar\chi$ (see (4.3)).

Step (v): Derivation of (6.32). We shall only outline the argument, since in most respects it is close to that in steps (i) and (ii). First we treat (6.32) for $j = 1$. Put

$$\begin{aligned} s_1(x, y) &= w(x^{\mathrm{c}}) \left\{ \frac{\mu_G(x)}{\mu_m(x)^2} - 2\, \frac{\mu_f(x, y)}{\mu_m(x)} \right\} = n^{-1}\, s_{11}(x, y) + s_{12}(x, y)\,, \\ s_{11}(x, y) &= w(x^{\mathrm{c}}) \left\{ \tau(x) \int \frac{\bar\mu_{f2}(\bar x, u)}{\bar\mu_m(\bar x)^2}\, du - \int \frac{\bar\mu_f(x, u)^2}{\bar\mu_m(\bar x)^2}\, du \right\}, \\ s_{12}(x, y) &= w(x^{\mathrm{c}}) \left\{ \int \frac{\bar\mu_f(\bar x, u)^2}{\bar\mu_m(\bar x)^2}\, du - 2\, \frac{\bar\mu_f(\bar x, y)}{\bar\mu_m(\bar x)} \right\} \end{aligned}$$

$$= w(x^{\rm c}) \biggl\{ \int \bar\mu_g(u|\bar x)^2 \, du - 2 \, \bar\mu_g(y|\bar x) \biggr\} ,$$

$$s_{13}(x,y) = w(x^{\rm c}) \biggl\{ \int \bar g(u|\bar x)^2 \, du - 2 \, \bar g(y|\bar x) \biggr\} , \quad s_{14}(x,y) = s_{12}(x,y) - s_{13}(x,y) \, .$$

In this notation, $S_1 = n^{-1} \sum_i s_1(X_i, Y_i)$.

Using Rosenthal's inequality it may be proved that, for all integers $k \ge 1$,

$$E\Bigl(n^{-1} \sum_{i=1}^n \bigl[s_{11}(X_i, Y_i) - E\{ s_{11}(X,Y) \} \bigr] \Bigr)^{2k} \le \text{const.} \, (nH)^{-k} \, ,$$

$$E\Bigl(n^{-1} \sum_{i=1}^n \bigl[s_{14}(X_i, Y_i) - E\{ s_{14}(X,Y) \} \bigr] \Bigr)^{2k} \le \text{const.} \, \bigl\{ (D/n)^k + n^{1-2k} \bigr\} \, ,$$

uniformly in $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ satisfying (6.4). Therefore, defining

$$Q_1 = n^{-1} \sum_{i=1}^n \bigl[s_{13}(X_i, Y_i) - E\{ s_{13}(X,Y) \} \bigr] \, ,$$

we have:

$$\begin{aligned} E(\Delta_1 - Q_1)^{2k} = E(S_1 - ES_1 - Q_1)^{2k} &\le \text{const.} \, \bigl\{ (n^3 H)^{-k} + (D/n)^k + n^{1-2k} \bigr\} \\ &\le \text{const.} \, \bigl[n^{-\delta} \bigl\{ (nH)^{-1} + D \bigr\} \bigr]^{2k} \, , \end{aligned} \tag{6.41}$$

where the last inequality holds for any sufficiently small $\delta > 0$, and δ does not depend on k.

By (6.41) and Markov's inequality,

$$\sup P \Bigl[\, \Bigl| \Delta_1(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) - Q_1 \, \Bigr| > n^{-\delta} \bigl\{ (nH)^{-1} + D \bigr\} \, \Bigr] = O\bigl(n^{-C} \bigr) \, , \tag{6.42}$$

for some $\delta > 0$ (smaller than that in (6.41)) and all $C > 0$, where the supremum has the same interpretation as that at (6.32). The supremum may be moved inside the probability in (6.42), by first noting that this is possible if the supremum is over a set $\mathcal{S}_n$ of bandwidth vectors that contains no more than polynomially many elements (as a function of n), and then observing that (since K and L are Hölder continuous) a sufficiently fine mesh of this type is adequate to transfer the supremum (already inside the probability) from $\mathcal{S}_n$ to the set of all vectors of smoothing parameters prescribed by (6.4). This completes the proof of (6.32) in the case $j = 1$.

Next we treat (6.32) in the case $j = 2$. It is clear from the definition of Δ_2 at (6.30) that it suffices to prove the version of (6.32) in which Δ_j is replaced by Δ_{2j}

for $1 \le j \le 4$, where $\Delta_{2j} = n^{-1} \sum_i s_{2j}(X_i, Y_i)$,

$$s_{21}(x,y) = s_{21}(x) = \frac{w(x^{\mathrm{c}})\,\Delta_{G,-i}(x)}{\mu_m(x)^2}\,, \quad s_{22}(x,y) = \frac{w(x^{\mathrm{c}})\,\Delta_{f,-i}(x,y)}{\mu_m(x)}\,,$$
$$s_{23}(x,y) = \frac{w(x^{\mathrm{c}})\,\mu_G(x)\,\Delta_{m,-i}(x)}{\mu_m(x)^3}\,, \qquad s_{24}(x,y) = \frac{w(x^{\mathrm{c}})\,\mu_f(x,y)\,\Delta_{m,-i}(x)}{\mu_m(x)^2}\,.$$

For the sake of brevity we shall discuss only the cases $j = 1, 2$, treating the second first.

We may write

$$\Delta_{22} = \frac{1}{n(n-1)} \sum\sum_{i_1 \ne i_2} \psi(X_{i_1}, X_{i_2}, X_{i_2}, Y_{i_2})\,,$$

where

$$\psi(x_1, y_1, x_2, y_2) = \frac{w(x_1^{\mathrm{c}})}{\mu_m(x_1)} \left\{K(x_1, x_2)\, L(y_1, y_2) - EK(x_1, X)\, L(y_1, Y)\right\}$$

and satisfies $E\{\psi(x, y, X, Y)\} = 0$. Put $\psi_1(x, y) = E\{\psi(X, Y, x, y)\}$, let ψ_2 denote the function that results from ψ_1 if we take

$$h = h_1 = \ldots = h_{p_1} = \lambda_1 = \ldots = \lambda_{q_1} = 0\,,$$
$$h_{p_1+1} = \ldots = h_p = \infty\,, \quad \lambda_j = (r_j - 1)/r_j \text{ for } q_1 + 1 \le j \le q\,,$$

and define

$$\Delta_{221} = \frac{1}{n(n-1)} \sum\sum_{i_1 \ne i_2} \left\{\psi(X_{i_1}, Y_{i_1}, X_{i_2}, Y_{i_2}) - \psi_1(X_{i_2}, Y_{i_2})\right\},$$
$$\Delta_{222} = \frac{1}{n} \sum_{i=1}^n \psi_1(X_i, Y_i)\,, \quad Q = \frac{1}{n} \sum_{i=1}^n \psi_2(X_i, Y_i)\,.$$

Note that Q does not depend on smoothing parameters.

The argument leading to (6.41) may be used to show that

$$E(\Delta_{222} - Q)^{2k} \le \text{const.} \left\{(D/n)^k + n^{1-2k}\right\} \le \text{const.} \left[n^{-\delta} \left\{(nH)^{-1} + D\right\}\right]^{2k}. \quad (6.43)$$

The martingale methods employed in steps (i) and (ii) may be employed to prove that for some $\delta > 0$ we have for all sufficiently large k,

$$E\big(\Delta_{221}^{2k}\big) \le C_k(\delta)\, n^{-k\delta}\, (nH)^{-2k}\,. \quad (6.44)$$

The heuristic argument behind (6.44), analogous to that at (6.12), is:

$$E\big(\Delta_{221}^{2k}\big) \le \text{const.}\,(\text{var}\,\Delta_{221})^{k-c_1} \le \text{const.}\,\big(n^2H\big)^{c_2-k} = \text{const.}\,n^{2c_2}\,H^{k-c_2}\,(nH)^{-2k}\,,$$

where $c_1, c_2 > 0$. Combining (6.43) and (6.44) we deduce that (6.42) holds if we replace (Δ_1, Q_1) on the left-hand side by (Δ_{22}, Q). Hence, (6.32) holds if we replace (Δ_1, Q_1) there by (Δ_{22}, Q).

A similar argument leads to the same result if (Δ_1, Q_1) at (6.32) is replaced by (Δ_{21}, Q) for another random variable Q not depending on the smoothing parameters. To appreciate why, write

$$\Delta_{21} = \frac{1}{n(n-1)(n-2)} \sum\sum\sum_{i_1,i_1,i_3 \text{ distinct}} \psi(X_{i_1}, X_{i_2}, Y_{i_2}, X_{i_3}, Y_{i_3})\,, \tag{6.45}$$

where

$$\psi(x, x_1, y_1, x_2, y_2) = \frac{w(x^{\mathrm{c}})}{\mu_m(x)^2} \bigg[K(x, x_1)\,K(x, x_2) \int L(y, y_1)\,L(y, y_2)\,dy \\ - E\bigg\{ K(x, X_1)\,K(x, X_2) \int L(y, Y_1)\,L(y, Y_2)\,dy \bigg\} \bigg]$$

and satisfies $E\{\psi(x, X_1, Y_1, X_2, Y_2)\} = 0$. Define

$$\phi_1 = \psi - \psi_1 - \psi_2 - \psi_3 + \psi_4 + \psi_5\,, \quad \phi_2 = \psi_1 - \psi_4 - \psi_5\,, \quad \phi_3 = \psi_2 - \psi_5\,, \quad \phi_4 = \psi_3 - \psi_4\,,$$

where $\psi_1(x_1, y_1, x_2, y_2) = E\{\psi(X_{i_1}, x_1, y_1, x_2, y_2)\}$,

$$\psi_2(x, x_2, y_2) = E\{\psi(x, X_{i_2}, Y_{i_2}, x_2, y_2)\}\,, \quad \psi_3(x, x_1, y_1) = E\{\psi(x, x_1, y_1, X_{i_3}, Y_{i_3})\}\,,$$
$$\psi_4(x_1, y_1) = E\{\psi(X_{i_1}, x_1, y_1, X_{i_3}, Y_{i_3})\}\,, \quad \psi_5(x_2, y_2) = E\{\psi(X_{i_1}, X_{i_2}, Y_{i_2}, x_2, y_2)\}\,.$$

Note that

$$\psi = \phi_1 + \phi_2 + \phi_3 + \phi_4 + \psi_4 + \psi_5\,. \tag{6.46}$$

Let Δ_{2j}, for $2 \le j \le 7$, denote the versions of Δ_{21} at (6.45) that arise if we replace the function ψ there by ϕ_1, ϕ_2, ϕ_3, ϕ_4, ψ_4 and ψ_5, respectively. Then by (6.46),

$$\Delta_{21} = \sum_{j=2}^{7} \Delta_{2j}\,. \tag{6.47}$$

If, in the functions ϕ_j, we replace (x, x_1, y_1, x_2, y_2) by $(X_{i_1}, X_{i_2}, Y_{i_2}, X_{i_3}, Y_{i_3})$, where i_1, i_2 and i_3 are distinct, then ϕ_1 is a function of three independent quantities,

and its expected value conditional on any one or two of them equals zero; ϕ_2, ϕ_3 and ϕ_4 are functions of two independent quantities, and their expected values conditional on any one of them equal zero; and ψ_4 and ψ_5 are functions of just one quantity, and have expected value zero.

In view of these properties, Δ_{22} is a degenerate U-statistic of order 3; Δ_{23}, Δ_{24} and Δ_{25} are degenerate U-statistics of order 2; and Δ_{26} and Δ_{27} are sums of independent random variables. After subtracting from the latter their respective counterparts in which all smoothing parameters are set equal to their asymptotic limits, one obtains a quantity whose $2k$th moments enjoy the bound at (6.43). The $2k$th moments of Δ_{23}, Δ_{24} and Δ_{25} admit the bound at (6.44), and a similar argument shows that the same bound applies to the $2k$th moment of Δ_{22}. Combining these properties, and noting (6.47), we deduce that for a random variable Q not depending on the smoothing parameters, (6.42) holds if we replace (Δ_1, Q_1) on the left-hand side by (Δ_{21}, Q). Hence, (6.32) holds if we replace (Δ_1, Q_1) there by (Δ_{21}, Q).

6. PROOF OF THEOREM 4.2

We begin with a stochastic approximation to $\hat{f}/\widehat{m}$, as follows. Defining $\mu_m = E(\widehat{m})$, $\mu_f = E(\hat{f})$, $\Delta_m = \widehat{m} - \mu_m$ and $\Delta_f = \hat{f} - \mu_f$, it may be proved that for some $\delta > 0$ and all $C > 0$,

$$P\biggl(\sup \biggl| \frac{\hat{f}(x,y)\, w(x^{\mathrm{c}})}{\widehat{m}(x)} - \biggl[\frac{\mu_f(x,y)\, w(x^{\mathrm{c}})}{\mu_m(x)} \biggl\{ 1 - \frac{\Delta_m(x)}{\mu_m(x)} \biggr\} + \frac{\Delta_f(x,y)\, w(x^{\mathrm{c}})}{\mu_m(x)} \biggr] \biggr| > n^{-\delta}\, (nH)^{-1} \biggr) = O\bigl(n^{-C}\bigr)\,, \tag{6.1}$$

where the supremum is of the stated order uniformly in $x, y, h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ prescribed by:

> $1 \le i \le n$ and (x,y) such that $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$, with x^{d} in the support of the distribution of X^{d}; $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ such that $h \le n^{-\epsilon}$ and $n^{\epsilon-1} \le H \le n^{-\epsilon}$; and uniformly in $\lambda_j \in \Lambda_j$ for $1 \le j \le q$;

for arbitrary but fixed $\epsilon \in (0, \frac{1}{2})$.

Let η_n denote any positive sequence decreasing to zero, and let $\mathcal{S}_n$ be the set of all vectors $v = (h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ of smoothing parameters that satisfy the properties: $h = n^{-1/(p_1+5)}\, a_0$, $h_j = n^{-1/(p_1+5)}\, a_j$ for $1 \le j \le p_1$, $h_j \ge \eta_n^{-1}$ for

$p_1 + 1 \le j \le p$, $\lambda_j = n^{-2/(p_1+5)}\, b_j$ for $1 \le j \le q_1$, and $\lambda_j = (r_j - 1)/r_j - \delta_j$ for $q_1 + 1 \le j \le q$, where $a_j \in [a_j^0 - \eta_n, a_j^0 + \eta_n]$, $b_j \in [\max(0, b_j^0 - \eta_n), b_j^0 + \eta_n]$ and $\delta_j \in [0, \eta_n]$.

Result (6.1) implies that

$$\hat{g}(y|x) = \frac{\mu_f(x,y)}{\mu_m(x)} \left\{ 1 - \frac{\Delta_m(x)}{\mu_m(x)} \right\} - \frac{\Delta_f(x,y)}{\mu_m(x)} + o_p\big(n^{-2/(p_1+5)}\big)\,, \tag{6.2}$$

uniformly in $v \in \mathcal{S}_n$. Throughout the present paragraph, let $v = v_n$ denote any deterministic smoothing parameter sequence in $\mathcal{S}_n$. Property (4.10) is a consequence of (6.2) and the results:

$$\frac{\mu_f(x,y)}{\mu_m(x)} = \frac{\bar{\mu}_f(\bar{x},y)}{\bar{\mu}_m(\bar{x})} = \bar{g}(y|\bar{x}) + n^{-2/(p_1+5)}\, \beta(\bar{x},y) + o\big(n^{-2/(p_1+5)}\big)\,, \tag{6.3}$$

$$\frac{n^{2/5}}{\nu_1(x)\, \bar{m}(\bar{x})} \left\{ \bar{g}(\bar{x},y)\, \Delta_m(x) + \Delta_f(x,y) \right\} \to \mathrm{N}\big\{0, \sigma(\bar{x},y)^2\big\}\,, \tag{6.4}$$

$$\left| \frac{\Delta_m(x)}{\nu_1(x)} \right| + \left| \frac{\Delta_f(x,y)}{\nu_1(x)} \right| = O_p\big(n^{-2/(p_1+5)}\big)\,, \tag{6.5}$$

where the convergence in (6.4) is in distribution. Property (6.3) follows by elementary calculus, (6.4) by a central limit theorem for sums of independent random variables, and (6.5) by calculating the mean square of the left-hand side and showing that it equals $O(n^{-4/(p_1+5)})$.

We shall apply the superscript * to a quantity to denote the value it takes for any particular, deterministic, but otherwise arbitrary, $v \in \mathcal{S}_n$. In order to extend (4.10) to the case where the smoothing parameter sequence is stochastic, and in particular obtained by cross-validation, it suffices to show that

$$\sup_{v \in \mathcal{S}_n} \left| \frac{\bar{\mu}_f(\bar{x},y)}{\bar{\mu}_m(\bar{x})} - \frac{\bar{\mu}_f^*(\bar{x},y)}{\bar{\mu}_m^*(\bar{x})} \right| = o\big(n^{-2/(p_1+5)}\big)\,, \tag{6.6}$$

$$\sup_{v \in \mathcal{S}_n} \left| \frac{\Delta_f^*(x,y)}{\nu_1^*(x)} - \frac{\Delta_f(x,y)}{\nu_1(x)} \right| = o_p\big(n^{-2/(p_1+5)}\big)\,, \tag{6.7}$$

$$\sup_{v \in \mathcal{S}_n} \left| \frac{\Delta_m^*(x)}{\nu_1^*(x)} - \frac{\Delta_m(x)}{\nu_1(x)} \right| = o_p\big(n^{-2/(p_1+5)}\big)\,. \tag{6.8}$$

Result (6.6) follows by elementary calculus, so it suffices to derive (6.7) and (6.8). We shall confine attention to (6.7).

Write $\Delta_f^\dagger$ for the version of Δ_f that is obtained by omitting the last p_2 components of the X_i^{c}'s and the last q_2 components of the X_i^{d}'s. That is, defining

$$A_i(x,y) = \left\{ \prod_{j=1}^{p_1} \frac{1}{h_j}\, K\left(\frac{x_j^{\mathrm{c}} - X_{ij}^{\mathrm{c}}}{h_j} \right) \right\}$$

$$\times \left\{ \prod_{j=1}^{q} \left(\frac{\lambda_j}{r_j - 1} \right)^{N_{ij}(x)} (1 - \lambda_j)^{1 - N_{ij}(x)} \right\} \frac{1}{h} L\left(\frac{y - Y_i}{h} \right),$$

we put $\Delta_f^\dagger = n^{-1} \sum_i (A_i - EA_i)$. Let

$$\Delta_f^\#(x, y) = \frac{\Delta_f(x, y)}{\nu_1(x)} - \Delta_f^\dagger(x, y)\,.$$

Elementary moment calculations show that $E(\Delta_f^\#)^2 = o(n^{-4/(p_1+5)})$, uniformly in smoothing parameters $v \in \mathcal{S}_n$. Using properties of rates of convergence in invariance principles for multivariate empirical processes (see e.g. Rio, 1996), these results may be generalised by showing that the normalised stochastic process, indexed by $v \in \mathcal{S}$, converges to zero uniformly in smoothing parameters in $\mathcal{S}_n$:

$$n^{2/(p_1+5)} \sup_{v \in \mathcal{S}_n} \left| \Delta_f^\#(x, y) \right| \to 0$$

in probability.

Therefore, in order to establish (6.7) it suffices to prove that

$$\sup_{v \in \mathcal{S}_n} \left| \Delta_f^\dagger(x, y) - \Delta_f^{\dagger *}(x, y) \right| = o\big(n^{-2/(p_1+5)}\big)\,, \tag{6.9}$$

where $\Delta_f^{\dagger *}(x, y)$ denotes the version of $\Delta_f^\dagger(x, y)$ computed for any particular value of the smoothing parameter vector v. (Note that neither $\Delta_f^{\dagger *}(x, y)$ nor $\Delta_f^\dagger(x, y)$ depends on the last p_2 h_j's or the last q_2 λ_j's.) Simple moment calculations show that $E\{\Delta_f^\dagger(x, y) - \Delta_f^{\dagger *}(x, y)\}^2 = o(n^{-4/(p_1+5)})$ uniformly in $v \in \mathcal{S}_n$, and this result may again be extended, to (6.9), using properties of invariance principles for multivariate empirical processes. Therefore (6.7) holds, completing the proof of the theorem.

Acknowledgements. We are grateful to an Associate Editor and two referees for constructive criticism.

REFERENCES

AITCHISON, J. AND AITKEN, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–420.

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

BURMAN, P. (1987). Smoothing sparse contingency tables. *Sankhyā* Ser. A **49**, 24–36.

FAN, J. AND YIM, T.H. (2003). A data-driven method for estimating conditional densities. Manuscript.

FRIEDMAN, J.H. AND STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817–823.

FRIEDMAN, J.H., STUETZLE, W. AND SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79**, 599–608. 62G05

HABBEMA, J.D.F., HERMANS, J. AND VAN DEN BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 1974*, ed. G. Bruckmann, pp. 101–110. Physica Verlag, Vienna.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66 (2)**, 315-331.

HALL, P. (1983a). Orthogonal series methods for both qualitative and quantitative data. *Ann. Statist.* **11**, 1004–1007.

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

HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Multivariate Analysis VI*, ed. P.R. Krishnaiah, 289–309. North-Holland, Amsterdam.

HALL, P. AND HEYDE, C.C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York.

HALL, P. AND MARRON, J.S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theor. Related Fields* **74**, 567–581.

HALL, P. AND TITTERINGTON, D.M. (1987). On smoothing sparse multinomial data. *Austral. J. Statist.* **29**, 19–37.

HIRANO, K. AND IMBENS, G. AND RIDDER, G. (2002). Efficient estimation of average treatment effects using the estimated propensity Score. *Econometrica*, to appear.

HOROWITZ, J. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, **69 (2)**, 499–513.

HSIAO, C. AND TAHMISCIOGLU, L. (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association*, **92**, 455–465.

JONES, M.C. AND SIBSON, R. (1987). What is projection pursuit? (With discussion.) *J. Roy. Statist. Soc.* Ser. A **150**, 1–36.

KLEIN R. AND SPADY, R. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–422.

LEWBEL, A. AND LINTON, O. (2002). Nonparametric censored and truncated regression. *Econometrica*, **70 (2)**, 765–779.

LI, Q. AND RACINE, J. (2003). Nonparametric estimation of joint distribution with mixed continuous and categorical data. *J. Multivar. Anal.*, **86**, 266–292.

POWELL, J.L., STOCK, J.H. AND STOKER, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403–1430.

RIO, E. (1996). Vitesses de convergence dans le principe d'invariance faible pour la fonction de répartition empirique multivariée. *C. R. Acad. Sci. Paris Sér. I Math.* **322**, 169–172.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

SAIN, S.R. (2001). Bias reduction and elimination with kernel estimators. *Commun. Statist.–Theor. Meth.* **30**, 1869–1888.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. Springer Verlag, New York.

STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285–1297.

TUTZ, G. (1991). Sequential models in categorical regression. *Comput. Statist. Data Anal.* **11**, 275–295.

WANG, M.C. AND VAN RYZIN, J.A. (1981). A class of smooth estimators for discrete distributions. *Biometrika* **68**, 301–309.