Cross-Validation and the Estimation of Conditional Probability Densities

Peter HALL, Jeff RACINE, and Qi LI

Many practical problems, especially some connected with forecasting, require nonparametric estimation of conditional densities from mixed data. For example, given an explanatory data vector $\mathbf{X}$ 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 $(\mathbf{X}, Y)$ for previous clients. Choosing appropriate smoothing parameters for this problem can be tricky, not in the 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 $\mathbf{X}$ are relevant to the problem of conditional inference. For example, if the jth component of $\mathbf{X}$ is independent of Y, then that component is irrelevant to estimating the density of Y given $\mathbf{X}$, and ideally should be dropped before conducting inference. In this article 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 toward 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 yields important information about which components are relevant; the relevant components are precisely those that 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.

KEY WORDS: 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.

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 $\mathbf{X}$ 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 $\mathbf{X}$ 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, because then conditional inference will be based on data with dimension that 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 $\mathbf{X}$ is p-variate and has a continuous distribution, then a conventional estimator of the density $g(y|\mathbf{x})$ of Y given $\mathbf{X} = \mathbf{x}$, using second-order kernels and a sample of size n, converges at rate $n^{-2/(p+5)}$ (assuming that 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 $\mathbf{X}$ (e.g., 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 $\mathbf{X}$ 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. In contrast, if we could identify components of $\mathbf{X}$ that were independent of Y, then we could delete these at the outset, leading to improvements in the accuracy with which the density of Y, given $\mathbf{X}$, 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 Sec. 5, based on two classic benchmark datasets.) In principle, this problem can be tackled by applying a battery of hypothesis tests before conducting inference. Tests for independence of individual components or 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 in the least because the components of $\mathbf{X}$ often will be of many different types—e.g., continuous, unordered discrete, ordered discrete—all in the same vector. We 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 $\mathbf{X}$ are continuous; this will simplify exposition. Construction of the cross-validation criterion, say CV, is not trivial, but a certain weighted form of

Peter Hall is Professor of Statistics, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (E-mail: *peter.hall@anu.edu.au*). Jeff Racine is Associate Professor of Economics, Department of Economics and Center for Policy Research, Syracuse University, Syracuse, NY 13244. Qi Li is Hugh Roy Cullen Professor in Liberal Arts, Department of Economics, Texas A&M University, College Station, TX 77843. The authors thank an associate editor and two referees for constructive criticism.

Journal of the American Statistical Association
December 2004, Vol. 99, No. 468, Theory and Methods
DOI 10.1198/016214504000000548

it has an elementary form. If p_2 of the components of $\mathbf{X}$ are independent of Y and therefore irrelevant, and if the remaining p_1 components of $\mathbf{X}$ are relevant, then the empirical bandwidths that minimize CV will demonstrate markedly dichotomous behavior: those that correspond to irrelevant components will diverge to infinity as the sample size increases, and 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 input 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 behavior is observed when one or more of the components of $\mathbf{X}$ 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 estimating the conditional density of Y given $\mathbf{X}$. 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 with both continuous and discrete components. For simplicity, in our theoretical work we 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 characterized, 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 only address the setting in which a different smoothing parameter is used for each component. Similar results are obtained 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 $\mathbf{X}$ 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, characterizing 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 use 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 is no 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 formulas are available. In additional, plug-in rules, even after adaptation to mixed data, require choosing of "pilot" smoothing parameters, and it is not clear how to best make that selection for the continuous and discrete variables involved. As we discuss later, cross-validation avoids these problems and has the additional virtue of separating variables into relevant and irrelevant categories.

Our method can readily be generalized 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 generalized additive model with 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). The cross-validation based smoothing method presented in this article can be used to generalize 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 this approach, let us again assume for simplicity that all of the components of $\mathbf{X}$ are continuous. We might say that $\mathbf{X} = (\mathbf{X}^{[1]}, \mathbf{X}^{[2]})$ represents a decomposition of $\mathbf{X}$ into relevant and irrelevant parts $\mathbf{X}^{[1]}$ and $\mathbf{X}^{[2]}$, if Y and $\mathbf{X}^{[2]}$ are independent conditional on $\mathbf{X}^{[1]}$. Although 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 $j = 1, 2$, $\mathbf{X} = (X_1, X_2)$, $\epsilon > 0$, 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 upon, a practical assessment of "relevance" based on conditional independence is likely to suggest that either X_1 or X_2 (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 $\mathbf{X}$ into relevant and irrelevant parts. On the other hand, if an unconditional view of independence is taken, as suggested in the present article, 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 earlier 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. Alternative techniques might be developed for dealing with the ambiguity described in the previous paragraph.

Section 2 introduces our cross-validation algorithm, and Section 3 develops properties of mean squared error (MSE) and optimal smoothing parameters when no irrelevant components are present. The results given there set theoretical benchmarks for the performance of bandwidth selectors after irrelevant components have been removed. Section 4 shows that cross-validation attains these benchmarks. There we give concise, mathematical definitions of what we mean by "relevant" and "irrelevant" components. Section 5 gives numerical illustrations of the performance of cross-validation in removing irrelevant components and conducting adaptive inference. 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 with large numbers of discrete cells relative to their sample sizes (thus the conventional frequency method is infeasible for both datasets). We show that our proposed estimator smooths 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 (1983a, 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, sec. 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 the densities 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 (2004), who discussed novel methods for conditional density estimation.

2. METHODOLOGY FOR CROSS–VALIDATION

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

$$\mathrm{ISE} = \int \{\hat{g}(y|\mathbf{x}) - g(y|\mathbf{x})\}^2 m(\mathbf{x})\, dW(\mathbf{x})\, dy, \tag{1}$$

where $dW(\mathbf{x})$ denotes the infinitesimal element of a measure.

The presence of $dW(\mathbf{x})$ at (1) serves only to avoid difficulties caused by dividing by 0, or by numbers close to 0, in the ratio $\hat{f}(\mathbf{x}, y)/\widehat{m}(\mathbf{x})$. This is usually a problem only for the continuous components of $\mathbf{x}$. Therefore, if $\mathbf{X}$ denotes a generic $\mathbf{X}_i$, if $\mathbf{X} = (\mathbf{X}^{\mathrm{c}}, \mathbf{X}^{\mathrm{d}})$ represents a division of $\mathbf{X}$ into continuous and discrete components, and if $\mathbf{x} = (\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}})$ is the corresponding division of $\mathbf{x}$, then we take $dW(\mathbf{x}) = w(\mathbf{x}^{\mathrm{c}})\, d\mathbf{x}^{\mathrm{c}}\, dV(\mathbf{x}^{\mathrm{d}})$, where $dV(\mathbf{x}^{\mathrm{d}})$ denotes the infinitesimal element of the Dirac delta measure that places unit mass at each atom of $\mathbf{x}^{\mathrm{d}}$. In this notation, (1) can be written equivalently as

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

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

We assume that $\mathbf{X}^{\mathrm{c}}$ is p-variate and $\mathbf{X}^{\mathrm{d}}$ q-variate. 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 et al. (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 consequently, our theoretical results in Section 3 will not strictly apply after dimension reduction.

Our estimators of f and m will be of the kernel type,

$$\begin{aligned} \hat{f}(\mathbf{x}, y) &= n^{-1} \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i) L(y, Y_i) \qquad \text{and} \\ \widehat{m}(\mathbf{x}) &= n^{-1} \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i), \end{aligned} \tag{3}$$

where K and L are nonnegative, generalized kernels. As (3) suggests, we use the same vector of smoothing parameters (one for each component) when treating the explanatory variables $\mathbf{X}_i$, regardless of whether we are addressing the numerator or the denominator of the estimator $\hat{g}(y|\mathbf{x}) = \hat{f}(\mathbf{x}, y)/\widehat{m}(\mathbf{x})$.

This "convention" guarantees that for each fixed $\mathbf{x}$ such that $\widehat{m}(\mathbf{x}) \neq 0$, $\hat{g}(\cdot|\mathbf{x})$ 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 $K(\mathbf{x}, \mathbf{X}_i)$. Reflecting the division $\mathbf{X} = (\mathbf{X}^{\mathrm{c}}, \mathbf{X}^{\mathrm{d}})$, write $\mathbf{X}_i = (\mathbf{X}_i^{\mathrm{c}}, \mathbf{X}_i^{\mathrm{d}})$, where $\mathbf{X}_i^{\mathrm{d}} = (X_{i1}^{\mathrm{d}}, \ldots, X_{iq}^{\mathrm{d}})$ and $\mathbf{X}_i^{\mathrm{c}} = (X_{i1}^{\mathrm{c}}, \ldots, X_{ip}^{\mathrm{c}})$ denote the discrete and continuous components of $\mathbf{X}_i$. (In particular, we no longer use the notation $\mathbf{X}_j^{\mathrm{c}}$ and $\mathbf{X}_j^{\mathrm{d}}$ for the jth components of $\mathbf{X}^{\mathrm{c}}$ and $\mathbf{X}^{\mathrm{d}}$.) We assume that X_{ij}^{d} takes the values $0, 1, \ldots, r_j - 1$. Put $\mathbf{x}^{\mathrm{c}} = (x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})$ and $\mathbf{x}^{\mathrm{d}} = (x_1^{\mathrm{d}}, \ldots, x_q^{\mathrm{d}})$, and define

$$K^{\mathrm{c}}(\mathbf{x}^{\mathrm{c}}, \mathbf{X}_i^{\mathrm{c}}) = \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x_j^{\mathrm{c}} - X_{ij}^{\mathrm{c}}}{h_j}\right),$$

where K is a traditional kernel function (i.e., a symmetric, univariate probability density) and

$$K^{\mathrm{d}}(\mathbf{x}^{\mathrm{d}}, \mathbf{X}_i^{\mathrm{d}}) = \prod_{j=1}^{q} \{\lambda_j/(r_j - 1)\}^{N_{ij}(\mathbf{x})} (1 - \lambda_j)^{1 - N_{ij}(\mathbf{x})}, \tag{4}$$

where $N_{ij}(\mathbf{x}) = I(X_{ij}^{\mathrm{d}} \neq x_j^{\mathrm{d}})$, depending on x_j^{d} alone, and I is the usual indicator function. In these formulas, $h_1, \ldots, h_p$ are bandwidths for the continuous components of $\mathbf{X}$ 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, $K^{\mathrm{d}}(\mathbf{x}^{\mathrm{d}}, \mathbf{X}_i^{\mathrm{d}})$ becomes unrelated to $(x_j^{\mathrm{d}}, X_{ij}^{\mathrm{d}})$ (i.e., the jth component of $\mathbf{x}^{\mathrm{d}}$ is completely smoothed out).

Formula (4) describes kernels that are appropriate for unordered categorical data (see, e.g., Aitchison and Aitken 1976). In the ordered case, alternative approaches can be used, using in effect near-neighbor weights (see, e.g., Wang and van Ryzin 1981; Burman 1987; Hall and Titterington 1987). In each case the kernel weights are intrinsically different from their continuum counterparts. In particular, for the weights defined in (4), in the asymptotic limit as each λ_j converges to 0, $K^{\mathrm{d}}(\mathbf{x}^{\mathrm{d}}, \mathbf{X}_i^{\mathrm{d}})$ converges to 1 if $\mathbf{X}_i^{\mathrm{d}} = \mathbf{x}^{\mathrm{d}}$ and to 0 otherwise. The resulting kernel-weighted estimator of the probability at $\mathbf{x}^{\mathrm{d}}$ converges to the naive cell-proportion (or maximum likelihood) estimator, which equals the proportion of the data for which $\mathbf{X}_i^{\mathrm{d}} = \mathbf{x}^{\mathrm{d}}$.

The generalized kernels, $K(\mathbf{x}, \mathbf{X}_i)$ and $L(y, Y_i)$, are given by

$$K(\mathbf{x}, \mathbf{X}_i) = K^{\mathrm{c}}(\mathbf{x}^{\mathrm{c}}, \mathbf{X}_i^{\mathrm{c}}) K^{\mathrm{d}}(\mathbf{x}^{\mathrm{d}}, \mathbf{X}_i^{\mathrm{d}}) \qquad \text{and}$$
$$L(y, Y_i) = \frac{1}{h} L\left(\frac{y - Y_i}{h}\right), \tag{5}$$

where L is another univariate kernel, possibly identical to K, and h is another bandwidth. The quantities at (5) are substituted into (3) to give $\hat{f}$ and $\widehat{m}$.

Expanding the right side of (1), we deduce that

$$\mathrm{ISE} = I_{1n} - 2I_{2n} + I_{3n}, \tag{6}$$

where

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

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

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

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

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

Thus, the cross-validation approximations $\hat{I}_{1n}$ and $\hat{I}_{2n}$, to I_{1n} and I_{2n}, are motivated as

$$\hat{I}_{1n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{G}_{-i}(\mathbf{X}_i) w(\mathbf{X}_i^{\mathrm{c}})}{\widehat{m}_{-i}(\mathbf{X}_i)^2}$$

and

$$\hat{I}_{2n} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{f}_{-i}(\mathbf{X}_i, Y_i) w(\mathbf{X}_i^{\mathrm{c}})}{\widehat{m}_{-i}(\mathbf{X}_i)},$$

where the subscript "$-i$" on a function of the data indicates that quantity is computed not from the n-sample, $\mathcal{Z} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, but rather from the $(n-1)$-sample, $\mathcal{Z} \backslash \{(\mathbf{X}_i, Y_i)\}$.

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 side of formula (6), but replaced by the approximations

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

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, then 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 are more than one of the values. Occasionally there is no local minimum in the range, in which case 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 that, in asymptotic terms, are optimal for minimizing the mean integrated squared error (MISE) defined by taking the expected value at (2),

$$\mathrm{MISE}(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) = \sum_{\mathbf{x}^{\mathrm{d}}} \int E\{\hat{g}(y|\mathbf{x}) - g(y|\mathbf{x})\}^2 m(\mathbf{x}) w(\mathbf{x}^{\mathrm{c}})\, d\mathbf{x}^{\mathrm{c}}\, dy. \tag{7}$$

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

Recall that $\mathbf{x} = (\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}})$, where $\mathbf{x}^{\mathrm{c}} = (x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})$ and $\mathbf{x}^{\mathrm{d}} = (x_1^{\mathrm{d}}, \ldots, x_q^{\mathrm{d}})$, and that w is a function of $\mathbf{x}^{\mathrm{c}}$. 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 $\mathbf{X}^{\mathrm{d}}$. We assume that:

> The densities f and m have two continuous derivatives as functions of $\mathbf{x}^{\mathrm{c}}$; w is continuous and nonnegative and has compact support, $m(\mathbf{x})$ is bounded away from 0 for $\mathbf{x} = (\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}}) \in \mathcal{S}^{\mathrm{c}} \times \mathcal{S}^{\mathrm{d}}$, and $\sup_{\mathbf{x} \in \mathcal{S}^{\mathrm{c}} \times \mathcal{S}^{\mathrm{d}}} f(\mathbf{x}, y)$ vanishes outside a compact set of values y. (8)

Let $f_{00}(\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}}, y)[f_{jj}(\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}}, y)]$ denote the second derivative of $f(\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}}, y)$ with respect to y [resp. x_j^{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 $I_j(\mathbf{u}^{\mathrm{d}}, \mathbf{x}^{\mathrm{d}})$ by

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

Note that $I_j(\mathbf{u}^{\mathrm{d}}, \mathbf{x}^{\mathrm{d}}) = 1$ if and only if $\mathbf{u}^{\mathrm{d}}$ and $\mathbf{x}^{\mathrm{d}}$ differ only at the jth component. Let $a_0, \ldots, a_p$ and $b_1, \ldots, b_q$ denote real numbers. We define a function of these quantities that represents the dominant term in an expansion of MISE [see (14) and (18)],

$$\begin{aligned}
&\chi(a_0, \ldots, a_p, b_1, \ldots, b_q) \\
&\quad = \sum_{\mathbf{x}^{\mathrm{d}}} \int \Bigg(\Bigg[\sum_{j=1}^{q} \frac{b_j}{r_j - 1} \sum_{\mathbf{u}^{\mathrm{d}}} I_j(\mathbf{u}^{\mathrm{d}}, x_j^{\mathrm{d}}) \\
&\qquad\qquad \times \left\{ f(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}}, y) - \frac{m(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}})}{m(\mathbf{x})} f(\mathbf{x}, y) \right\} \\
&\qquad\qquad + \frac{1}{2} \kappa_{L2} a_0^2 f_{00}(\mathbf{x}, y) \\
&\qquad\qquad + \frac{1}{2} \kappa_2 \sum_{j=1}^{p} a_j^2 \left\{ f_{jj}(\mathbf{x}, y) - \frac{m_{jj}(\mathbf{x})}{m(\mathbf{x})} f(\mathbf{x}, y) \right\} \Bigg]^2 \\
&\qquad\qquad + \frac{\kappa^p \kappa_L f(\mathbf{x}, y)}{a_0 \cdots a_p} \Bigg) \frac{w(\mathbf{x}^{\mathrm{c}})}{m(\mathbf{x})} \, d\mathbf{x}^{\mathrm{c}} \, dy, \qquad (9)
\end{aligned}$$

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

Write $a_0^0, \ldots, a_p^0$ and $b_1^0, \ldots, b_q^0$ for the values that minimize 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, considering that:

> The a_j^0's and b_j^0's are uniquely defined, and each is finite. (10)

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 $z_{p+1}, \ldots, z_{p+q}$:

$$\begin{aligned}
&\chi(z_0, \ldots, z_{p+q}) \\
&\quad = \int \left\{ \sum_{j=0}^{p+q} B_j(\mathbf{x}, y) z_j \right\}^2 d\mathbf{x} \, dy + \frac{c_0}{(z_0 \cdots z_p)^{1/2}} \\
&\quad = \mathbf{z}^{\mathrm{T}} \mathbf{A} \mathbf{z} + \frac{c_0}{(z_0 \cdots z_p)^{1/2}},
\end{aligned}$$

where $\mathbf{z} = (z_0, \ldots, z_{p+q})^{\mathrm{T}}$ and $\mathbf{A}$ is a $(p + q + 1) \times (p + q + 1)$ matrix. Then, if $\mathbf{A}$ 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 confine attention to

> $(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) \in [0, \eta]^{p+q+1}$, where $\eta = \eta_n$ denotes any positive sequence that satisfies $n^{\epsilon} \eta_n \to \infty$ for each $\epsilon > 0$. (11)

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 minimizing MSE, need not converge to 0.

Theorem 1. Assume (8) and (10), and that the smoothing parameters $h^0, h_1^0, \ldots, h_p^0, \lambda_1^0, \ldots, \lambda_q^0$ that minimize MISE are constrained by (11). Then

$$\begin{aligned}
&h^0 \sim a_0^0 n^{-1/(p+5)}, \\
&h_j^0 \sim a_j^0 n^{-1/(p+5)}, && \text{for } 1 \le j \le p, \qquad (12) \\
&\lambda_j^0 = b_j^0 n^{-2/(p+5)} + o\bigl(n^{-2/(p+5)}\bigr) && \text{for } 1 \le j \le q,
\end{aligned}$$

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

A key assumption in Theorem 1 is (10), which excludes some cases where components of $\mathbf{X}^{\mathrm{d}}$ or $\mathbf{X}^{\mathrm{c}}$ contain no effective information about Y. To appreciate this point, let $\mathbf{Z}^{\mathrm{c},-j}$ (resp., $\mathbf{Z}^{\mathrm{d},-j}$) denote the $(p + q)$-vector that arises after removing the jth component, $\mathbf{X}_j^{\mathrm{c}}$ (resp., $\mathbf{X}_j^{\mathrm{d}}$) of $\mathbf{X}^{\mathrm{c}}$ (resp. of $\mathbf{X}^{\mathrm{d}}$) from $\mathbf{Z} = (\mathbf{X}, Y)$. If $\mathbf{X}_j^{\mathrm{c}}$ and $\mathbf{Z}^{\mathrm{c},-j}$ were independent random variables, or if $\mathbf{X}_j^{\mathrm{d}}$ and $\mathbf{Z}^{\mathrm{d},-j}$ 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 $(\mathbf{X}_i, Y_i)$. We would instead delete the jth component of the continuous part of $\mathbf{X}_i$ or of the discrete part of $\mathbf{X}_i$.

In the first of these cases, $f_{jj} = (m_{jj}/m) f$, and so the term in a_j^2 vanishes from the right side of (9). As a result, $a_j^0 = \infty$, and so (10) is violated. It can be restored by excluding the jth component of $\mathbf{X}^{\mathrm{c}}$, as argued earlier. In the second case the quantity b_j in (9) can be absorbed into the other b_k's, and so the minimizer of χ is not uniquely defined. Therefore, (10) again fails, but it can be restored by dropping the jth component of $\mathbf{X}^{\mathrm{d}}$. In these instances it is fair to say that $\mathbf{X}_j^{\mathrm{c}}$ or $\mathbf{X}_j^{\mathrm{d}}$ is "completely irrelevant" for estimating g.

If $\mathbf{X}_j^{\mathrm{d}}$ were independent of Y but not independent of the other components of $\mathbf{X}$, 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 (10) was violated. This reflects the fact that the full-data vectors can produce estimators of m with lower MSE, and so may be beneficial. A similar argument applies if $\mathbf{X}_j^{\mathrm{c}}$ is independent of Y but not of other components of $\mathbf{X}$, although in this case a relatively standard kernel approach to estimation, advocated in (3), is not the best way to proceed.

3.2 Proof of Theorem 1

Using (13), it may be shown that ISE does not converge to 0 unless $Vh \to \infty$ as $n \to \infty$, where $V = n h_1 \cdots h_p$. Therefore, we may, without loss of generality, add to (11) the constraint that $Vh \ge t_n$, 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,

$$\begin{aligned}
&\hat{g}(y|\mathbf{x}) - g(y|\mathbf{x}) \\
&\qquad = \bigl\{ \hat{f}(\mathbf{x}, y) - \widehat{m}(\mathbf{x}) g(y|\mathbf{x}) + Q(\mathbf{x}, y) \bigr\} m(\mathbf{x})^{-1}, \qquad (13)
\end{aligned}$$

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

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

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

$$\mathrm{MISE} = \mathrm{MISE}_1 + o(\eta_1), \tag{14}$$

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

$$\begin{aligned}
&E\{\hat{f}(\mathbf{x}, y)\} \\
&\quad = \sum_{\mathbf{u}^{\mathrm{d}}} P(\mathbf{X}^{\mathrm{d}} = \mathbf{u}^{\mathrm{d}}) \left\{ \prod_{j=1}^{q} (1-\lambda_j) \rho_j^{u_j^{\mathrm{d}}(\mathbf{x})} \right\} \int \left\{ \prod_{j=1}^{p} K(z_j) \right\} L(v) \\
&\qquad \times \psi\big(x_1^{\mathrm{c}} - h_1 z_1, \ldots, \\
&\qquad\qquad x_p^{\mathrm{c}} - h_p z_p, y - hv | u_1^{\mathrm{d}}, \ldots, u_q^{\mathrm{d}}\big)\, dz_1 \cdots dz_p\, dv \\
&\qquad = f(\mathbf{x}, y) \\
&\qquad\quad + \sum_{j=1}^{q} \frac{\lambda_j}{r_j - 1} \left\{ \sum_{\mathbf{u}^{\mathrm{d}}} I_j(\mathbf{u}^{\mathrm{d}}, \mathbf{x}^{\mathrm{d}}) f(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}}, y) - f(\mathbf{x}, y) \right\} \\
&\qquad\quad + \frac{1}{2}\kappa_{L2} h^2 f_{00}(\mathbf{x}, y) + \frac{1}{2}\kappa_2 \sum_{j=1}^{p} h_j^2 f_{jj}(\mathbf{x}, y) + o(\eta_2),
\end{aligned}$$

where the remainders here and in (15)–(17) are of the stated size uniformly in $\mathbf{x}^{\mathrm{c}} \in \operatorname{supp} w$, in $\mathbf{x}^{\mathrm{d}}$ in the support of the distribution of $\mathbf{X}^{\mathrm{d}}$, and in y, as well as in smoothing-parameter vectors in $[0, \eta]^{p+q+1}$ satisfying $Vh \geq t_n$.

Similarly,

$$\begin{aligned}
&E\{\widehat{m}(\mathbf{x})\} \\
&\quad = m(x) + \sum_{j=1}^{q} \frac{\lambda_j}{r_j - 1} \left\{ \sum_{\mathbf{u}^{\mathrm{d}}} I_j(\mathbf{u}^{\mathrm{d}}, \mathbf{x}^{\mathrm{d}}) m(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}}) - m(\mathbf{x}) \right\} \\
&\qquad + \frac{1}{2}\kappa_2 \sum_{j=1}^{p} h_j^2 m_{jj}(\mathbf{x}) + o(\eta_2).
\end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned}
&E\{\hat{f}(\mathbf{x}, y)\} - E\{\widehat{m}(\mathbf{x})\} g(y|\mathbf{x}) \\
&\quad = \sum_{j=1}^{q} \frac{\lambda_j}{r_j - 1} \sum_{\mathbf{u}^{\mathrm{d}}} I_j(\mathbf{u}^{\mathrm{d}}, \mathbf{x}^{\mathrm{d}}) \\
&\qquad \times \left\{ f(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}}, y) - \frac{m(\mathbf{x}^{\mathrm{c}}, \mathbf{u}^{\mathrm{d}})}{m(\mathbf{x})} f(\mathbf{x}, y) \right\} \\
&\qquad + \frac{1}{2}\kappa_{L2} h^2 f_{00}(\mathbf{x}, y) \\
&\qquad + \frac{1}{2}\kappa_2 \sum_{j=1}^{p} h_j^2 \left\{ f_{jj}(\mathbf{x}, y) - \frac{m_{jj}(\mathbf{x})}{m(\mathbf{x})} f(\mathbf{x}, y) \right\} \\
&\qquad + o(\eta_2).
\end{aligned} \tag{16}$$

Note also that

$$\begin{aligned}
&n \operatorname{var}\{\hat{f}(\mathbf{x}, y) - \widehat{m}(\mathbf{x}) g(y|\mathbf{x})\} \\
&\quad = \operatorname{var}[K(\mathbf{x}, \mathbf{X}_i)\{L(y, Y_i) - g(y|\mathbf{x})\}] \\
&\quad = E\{K(\mathbf{x}, \mathbf{X}_i) L(y, Y_i)\}^2 + o(\eta_3) \\
&\quad = n\kappa^p \kappa_L f(\mathbf{x}, y) \eta_3 + o(\eta_3).
\end{aligned} \tag{17}$$

Combining (16) and (17), we deduce that

$$\begin{aligned}
&\mathrm{MISE}_1(h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q) \\
&\qquad = n^{-4/(p+1)} \chi(a_0, \ldots, a_p, b_1, \ldots, b_q) + o(\eta_1),
\end{aligned} \tag{18}$$

uniformly in smoothing-parameter vectors in $[0, \eta]^{p+q+1}$ satisfying $Vh \geq t_n$, 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 (14) and (18).

4. PROPERTIES OF CROSS–VALIDATION

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

> For integers $0 \leq p_1, p_2 \leq p$ and $0 \leq q_1, q_2 \leq q$ satisfying $p_1 + p_2 = p$ and $q_1 + q_2 = q$, the following is true: The vector $\mathbf{X}^{(1)}$ comprising the first p_1 components of $\mathbf{X}^{\mathrm{c}}$ and the first q_1 components of $\mathbf{X}^{\mathrm{d}}$ is stochastically independent of the vector $\mathbf{X}^{(2)}$ compristing the last p_2 components of $\mathbf{X}^{\mathrm{c}}$, and the last q_2 components of $\mathbf{X}^{\mathrm{d}}$, and $\mathbf{X}^{(2)}$ and Y are independent. (19)

A technical definition of "relevance" of the first p_1 components of $\mathbf{X}^{\mathrm{c}}$, and first q_1 components of $\mathbf{X}^{\mathrm{d}}$, is given in (21).

For the p_1 relevant continuous components and q_1 relevant discrete components of $\mathbf{X}$, we should impose the analog of assumption (10), asserting that the asymptotically optimal smoothing parameters are of conventional size. To this end, we introduce the following analog of the function χ at (9), now tailored to just the relevant components of $\mathbf{X}$:

$$\begin{aligned}
&\bar{\chi}(a_0, \ldots, a_{p_1}, b_1, \ldots, b_{q_1}) \\
&\quad = \sum_{\bar{\mathbf{x}}^{\mathrm{d}}} \int \Bigg(\Bigg[\sum_{j=1}^{q_1} \frac{b_j}{r_j - 1} \sum_{\bar{\mathbf{u}}^{\mathrm{d}}} I_j(\bar{\mathbf{u}}^{\mathrm{d}}, \bar{\mathbf{x}}^{\mathrm{d}}) \\
&\qquad\qquad \times \left\{ \bar{f}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{u}}^{\mathrm{d}}, y) - \frac{\bar{m}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{u}}^{\mathrm{d}})}{\bar{m}(\bar{\mathbf{x}})} \bar{f}(\bar{\mathbf{x}}, y) \right\} \\
&\qquad\qquad + \frac{1}{2}\kappa_{L2} a_0^2 \bar{f}_{00}(\bar{\mathbf{x}}, y) \\
&\qquad\qquad + \frac{1}{2}\kappa_2 \sum_{j=1}^{p_1} a_j^2 \left\{ \bar{f}_{jj}(\bar{\mathbf{x}}, y) - \frac{\bar{m}_{jj}(\bar{\mathbf{x}})}{\bar{m}(\bar{\mathbf{x}})} \bar{f}(\bar{\mathbf{x}}, y) \right\} \Bigg]^2 \\
&\qquad\qquad + \frac{\kappa^{p_1} \kappa_L \bar{f}(\bar{\mathbf{x}}, y)}{a_0 \cdots a_{p_1}} \Bigg) \bar{w}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{x}}^{\mathrm{d}})\, d\bar{\mathbf{x}}^{\mathrm{c}}\, dy,
\end{aligned} \tag{20}$$

where

$$\bar{w}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{x}}^{\mathrm{d}}) = \sum_{x^{\mathrm{d}}_{q_1+1}, \ldots, x^{\mathrm{d}}_{q}} \int \frac{w(\bar{\mathbf{x}}^{\mathrm{c}}, x^{\mathrm{c}}_{p_1+1}, \ldots, x^{\mathrm{c}}_{p})}{m(\bar{\mathbf{x}}^{\mathrm{c}}, x^{\mathrm{c}}_{p_1+1}, \ldots, x^{\mathrm{c}}_{p}, \bar{\mathbf{x}}^{\mathrm{d}}, x^{\mathrm{d}}_{q_1+1}, \ldots, x^{\mathrm{d}}_{q})} \times dx^{\mathrm{c}}_{p_1+1} \cdots dx^{\mathrm{c}}_{p},$$

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

The function $\bar{\chi}$ coincides exactly with χ in the case in which the last p_2 a_j's and last q_2 b_j's are deleted from the argument of χ and the weight $w(\mathbf{x}^{\mathrm{c}})/m(\mathbf{x})$ at (9) is replaced by $\bar{w}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{x}}^{\mathrm{d}})$. Of course, $\bar{w}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{x}}^{\mathrm{d}})$ is obtained from $w(\mathbf{x}^{\mathrm{c}})/m(\mathbf{x})$ on integrating (and summing) out the irrelevant components of $\mathbf{x}$ and is the weight function appropriate to the MISE that would be obtained at (7) if we were to drop all irrelevant components from the estimator $\hat{g}(y|\mathbf{x})$ appearing there. Therefore, we expect the optimal values of smoothing parameters in the present problem to be exactly those given by Theorem 1, but with (p, q) at (12) replaced by (p_1, q_1), and $h^0, h^0_1, \ldots, h^0_{p_1}, \lambda^0_1, \ldots, \lambda^0_{q_1}$ there chosen to minimize $\bar{\chi}$, defined in (20), rather than χ, given in (9). Theorem 2 later in this section shows that cross-validation selects smoothing parameters for the relevant components of $\mathbf{X}$ in precisely this asymptotically optimal manner.

Write $a^0_0, \ldots, a^0_{p_1}$ and $b^0_1, \ldots, b^0_{q_1}$ for the values that minimize $\bar{\chi}$, subject to each of them being nonnegative. The analog of condition (10) is that:

> The a^0_j's and b^0_j's are uniquely defined, and each is finite. (21)

To be able to detect the effect of relevant components of $\mathbf{X}$ on the conditional distribution of Y, within the domain to which we are constrained by the weight function w, we assume that

$$\sup_{\bar{\mathbf{x}} \in \operatorname{supp} \bar{w}} \sup_{y} \bar{g}(y|\bar{\mathbf{x}}) > 0. \tag{22}$$

The empirical observation that smoothing parameters chosen by cross-validation diverge to their upper extremes when the respective components of $\mathbf{X}$ 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 the 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 $\mathbf{X}$ are already uniformly distributed, then the effect of choosing the respective smoothing parameters may be comparatively small, and so different behavior can be expected. For the sake of simplicity, we consider 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,

$$R(\bar{\mathbf{x}}, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q) = \frac{E[\{\prod_{j=p_1+1}^{p} K(\frac{\bar{x}^{\mathrm{c}}_j - X^{\mathrm{c}}_{1j}}{h_j})^2\} \prod_{j=q_1+1}^{q} \{(1-\lambda_j)\rho_j^{N_{1j}(\bar{\mathbf{x}})}\}^2]}{(E[\{\prod_{j=p_1+1}^{p} K(\frac{\bar{x}^{\mathrm{c}}_j - X^{\mathrm{c}}_{1j}}{h_j})\} \prod_{j=q_1+1}^{q} \{(1-\lambda_j)\rho_j^{N_{1j}(\bar{\mathbf{x}})}\}])^2}.$$

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 that $K(0) > K(\delta)$ 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. But exceptions to this rule can arise if marginal distributions are uniform. We 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 $R(\bar{\mathbf{x}}, h_{p_1+1}, \ldots, h_p, \lambda_{q_1+1}, \ldots, \lambda_q) = 1$ for some $\bar{\mathbf{x}} \in \operatorname{supp} \bar{w}$, are $h_j = \infty$ for $p_1 + 1 \leq j \leq p$ and $\lambda_j = (r_j - 1)/r_j$ for $q_1 + 1 \leq j \leq q$. (23)

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 (11). By here that is not desirable, because we expect smoothing parameters corresponding to irrelevant components to be bounded away from 0. Constraint (11) would have us assume in advance that the bandwidths associated with relevant components converge to 0, yet for the sake of realism we do not wish those components to be identified to the experimenter. Therefore, we take a different tack, as follows.

Note that, in view of (19), the contributions of irrelevant components cancel from the ratio $\bar{\mu}_g(y|\bar{\mathbf{x}}) = E\{\hat{f}(\mathbf{x}, y)\}/E\{\widehat{m}(\mathbf{x})\}$. Let $\bar{g}$ denote the version of g when irrelevant components are dropped. We assume that:

> $\int dy \int_{\operatorname{supp} \bar{w}} \{\bar{\mu}_g(y|\bar{\mathbf{x}}) - \bar{g}(y|\bar{\mathbf{x}})\}^2 \, d\bar{\mathbf{x}}$, interpreted as a function of $h_1, \ldots, h_{p_1}$ and $\lambda_1, \ldots, \lambda_{q_1}$, vanishes if and only if all of those smoothing parameters vanish. (24)

Finally, we assume conventional conditions on the bandwidths and kernels. Define

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

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

> $0 < h \leq n^{-\epsilon}$; $n^{\epsilon-1} \leq H \leq n^{-\epsilon}$; $h_1 \cdots h_{p_1} h \geq n^{\epsilon-1}$; the kernels K and L are symmetric, compactly supported, Hölder-continuous probability densities; and $K(0) \neq 0$. (25)

We also impose the obvious requirement that the bandwidths that minimize 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 that

none of the bandwidths $h, h_1, \ldots, h_p$ is less than n^{-C_1} converges to 1 as $n \to \infty$. In concert with (25), this entails that for some $C_2 > 0$, the probability that none of the bandwidths $h, h_1, \ldots, h_{p_1}$ exceeds n^{C_2} converges to 1 as $n \to \infty$. Therefore, it may be supposed without loss of generality that

$$\text{for some } C > 0, \quad \min(h, h_1, \ldots, h_p) > n^{-C} \quad \text{and} \quad \max(h_1, \ldots, h_{p_1}) \le n^{C}. \tag{26}$$

We use this property in several places in our proofs, [e.g., in the second-to-the-last paragraph of step (a)], although we do not list it among the regularity conditions in Theorem 2.

We should comment on the extent to which (25) and (26) require knowledge of p_1, the number of relevant components of $\mathbf{X}$. The conditions do betray knowledge of whether p_1 is 0 or positive. However, we claim that under the following side condition, which does not depend on p_1, they hold whenever $1 \le p_1 \le p_2$:

$$\text{No } h_j, \text{ for } 1 \le j \le p_2, \text{ takes a value exceeding } D \log n, \text{ where } D > 0 \text{ is arbitrary; } h_1, \ldots, h_{p_2} \in [n^{\delta-1}, n^{-\delta}] \text{ for some } \delta \in (0, \tfrac{1}{2}); \text{ and } 1 \le p_1 \le p_2. \tag{27}$$

If this constraint holds, then (25) and (26) obtain for any $\epsilon \in (0, \delta)$ and $C \ge 1$, and all sufficiently large n, regardless of the value of p_1. Condition (27) is, of course, reasonable, in that it does not prevent $h_1, \ldots, h_{p_1}$ from taking appropriately small values, and at the same time it allows $h_{p_1}, \ldots, h_{p_2}$ to assume values that diverge to infinity.

Theorem 2. Assume (8), (19), and (22)–(25), and let $\hat{h}, \hat{h}_1, \ldots, \hat{h}_p, \hat{\lambda}_1, \ldots, \hat{\lambda}_q$ denote the smoothing parameters that minimize CV subject to the bandwidth constraints imposed in (25). Then, interpreting each convergence of a sequence of random variables as convergence in probability, we have

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

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

A proof of Theorem 2 is given in a longer version of this article, obtainable from the authors.

The conclusions of Theorem 2 may be summarized as follows. The smoothing parameters chosen by cross-validation, and corresponding to relevant components of the variables $\mathbf{X}_i$, have the properties of asymptotic optimality described by Theorem 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, for the method to make a mistake and, in effect, incorrectly remove relevant variables by choosing a too-large value of bandwidth. Results such as Theorem 2 state that the probability that this can happen converges to 0 as $n \to \infty$, but nevertheless there is always a nonzero probability that the method will do the wrong thing.

Next we discuss the performance of the empirical smoothing parameters when they are used to construct $\hat{g}$ at a point. We show that they produce an estimator that has the same first-order properties it would enjoy if the asymptotically optimal, deterministic parameters were used. The latter may be defined as: $h = n^{-1/(p_1+5)}a_0^0$, $h_j = n^{-1/(p_1+5)}a_j^0$ 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 $a_0^0, a_1^0, \ldots, a_{p_1}^0, b_1^0, \ldots, b_{q_1}^0$ minimize $\bar{\chi}$ defined at (20).

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

$$\hat{g}(y|\mathbf{x}) = g(y|\mathbf{x}) + n^{-2/(p_1+5)}\{\beta(\bar{\mathbf{x}}, y) + \sigma(\bar{\mathbf{x}}, y)Z_n(\mathbf{x}, y)\} + o_p\big(n^{-2/(p_1+5)}\big), \tag{28}$$

where the random variable $Z_n(\mathbf{x}, y)$ has the standard normal distribution,

$$\begin{aligned}
\beta(\bar{\mathbf{x}}, y) = {} & \sum_{j=1}^{q_1} \frac{b_j}{r_j - 1} \sum_{\bar{\mathbf{u}}^{\mathrm{d}}} I_j(\bar{\mathbf{u}}^{\mathrm{d}}, \bar{\mathbf{x}}^{\mathrm{d}})\\
& \times \left\{ \bar{g}(y|\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{u}}^{\mathrm{d}}) - \frac{\bar{m}(\bar{\mathbf{x}}^{\mathrm{c}}, \bar{\mathbf{u}}^{\mathrm{d}})}{\bar{m}(\bar{\mathbf{x}})} \bar{g}(y|\bar{\mathbf{x}}) \right\}\\
& + \frac{1}{2}\kappa_{L2} a_0^2 \bar{g}_{00}(y|\bar{\mathbf{x}})\\
& + \frac{1}{2}\kappa_2 \sum_{j=1}^{p_1} a_j^2 \left\{ \bar{g}_{jj}(y|\bar{\mathbf{x}}) - \frac{\bar{m}_{jj}(\bar{\mathbf{x}})}{\bar{m}(\bar{\mathbf{x}})} \bar{g}(y|\bar{\mathbf{x}}) \right\}
\end{aligned}$$

and

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

denote asymptotic bias and variance, and $\bar{g}_{jj}(y|\bar{\mathbf{x}})$ is the second derivative of $\bar{g}(y|\bar{\mathbf{x}})$ with respect to $\bar{x}_j^{\mathrm{c}}$. We give a proof of (28) as part of our derivation of Theorem 3 in Section 7. The theorem argues that (28) continues to hold if we choose the smoothing parameters empirically, by cross-validation.

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

Theorem 3. Assume the conditions imposed in Theorem 2; 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 $\mathbf{x} = (\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}}) \in \mathcal{S}^{\mathrm{c}} \times \mathcal{S}^{\mathrm{d}}$. Then (28) remains true, for the same functions β and σ, if $\hat{g}(y|\mathbf{x})$ 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, then we need to replace $L(y, Y_i) = h^{-1}L\{(y - Y_i)/h\}$, defined at (5), by $L(y, Y_i) = \lambda^{N_i(y)}(1 - \lambda)^{1 - N_i(y)}$, where $N_i(y) = I(Y_i \ne y)$. Then we need to modify Theorem 3 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 similarly 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 a modest Monte Carlo experiment designed to investigate the performance of the proposed estimator. We choose a popular setting, conditional binary prediction, and consider a latent variable probit data-generating process (DGP) with a mix of covariate types given by

$$\begin{aligned} Y_i^* &= \theta_1 Z_{1i} + \theta_2 Z_{2i} + \theta_3 X_i + \theta_4 Z_{1i} X_i \\ &\quad + \theta_5 Z_{2i} X_i + \theta_6 Z_{1i} Z_{2i} + \theta_7 Z_{1i} Z_{2i} X_i + \epsilon_i, \qquad (29) \\ Y_i &= 1 \quad \text{if} \quad Y_i^* > 0 \qquad \text{and} \qquad Y_i = 0 \quad \text{otherwise,} \end{aligned}$$

where the variables X_i, Z_{1i}, Z_{2i}, and ϵ_i, for $1 \le i \le n$ are totally independent; the X_i's are uniformly distributed on the interval $[0, 1]$; Z_{1i} and Z_{2i} take values only in the set $\mathcal{A} \equiv \{-2, -1, 0, 1, 2\}$; $\Pr(Z_{1i} = z) = \Pr(Z_{2i} = z) = .1, .4, .1, .35, .05$ as z ranges among the respective values in $\mathcal{A}$; and the ϵ_i's are normal $N(0, 1)$. We also treated the case where $\Pr(Z_1 = z) = \Pr(Z_2 = z) = 1/5$ for each $z \in \mathcal{A}$. The results that we obtained were qualitatively identical to those in the earlier setting and thus are not reported here.

We fix the sample size at $n_1 = 100$, and for each Monte Carlo replication compute the correct classification ratio (CCR) on independent evaluation data drawn from the same DGP ($n_2 = 1{,}000$) for each of three estimators, the proposed estimator, a parametric probit estimator, and the conventional (frequency) nonparametric estimator. The CCR is computed as the fraction of predicted Y equal to actual Y for the evaluation sample, and for each estimator we predict $\hat{Y} = 1$ if the estimated conditional probability $\widehat{\Pr}(Y = 1|z_1, z_2, x)$ exceeds .5 and $\hat{Y} = 0$ otherwise. We conduct the following three experiments:

1. For the DGP specified in (29), we set $\boldsymbol{\theta} = (1, 1, 1, 1, 1, 1, 1)^{\mathrm{T}}$, and let the parametric model be correctly specified (a probit model with index function given by $\beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 x_i + \beta_4 z_{1i} x_i + \beta_5 z_{2i} x_i + \beta_6 z_{1i} z_{2i} + \beta_7 z_{1i} z_{2i} x_i$).
2. We set $\boldsymbol{\theta} = (0, 0, 1, 0, 0, 0, 0)^{\mathrm{T}}$ and let both the parametric and the nonparametric models be overspecified [a probit with index function $\beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 x_i + \beta_4 z_{1i} x_i + \beta_5 z_{2i} x_i + \beta_6 z_{1i} z_{2i} + \beta_7 z_{1i} z_{2i} x_i$ and a function $g(y_i|x_i, z_{1i}, z_{2i})$ subject only to smoothness constraints].
3. We set $\boldsymbol{\theta} = (1, 1, 1, 1, 1, 1, 1)^{\mathrm{T}}$ and let the parametric model be underspecified (a probit with index function $\beta_1 z_{1i} + \beta_2 z_{2i} + \beta_3 x_i$).

We report the median CCR on the independent evaluation data for each model over the 1,000 Monte Carlo replications along with the 5th and 95th percentiles, and summarize these results in Tables 1–3.

We note that, compared with the correctly specified parametric model (Table 1), the proposed estimator exhibits a $< 5\%$ predictive efficiency loss. For the overspecified model (Table 2), however, the proposed estimator exhibits a 4% predictive efficiency gain. Although both the parametric and nonparametric models use the same information in this instance, and the estimator based on the parametric model has a faster rate of convergence, the cross-validated kernel estimator effectively removes the irrelevant variables from the resulting estimator, producing finite-sample efficiency gains, whereas the presence of irrelevant variables contributes additional noise to a parametric model. The median cross-validated bandwidths for the irrelevant variables Z_1 and Z_2 over the 1,000 Monte Carlo replications were only slightly below their maximal upper bound values of .8 ($\hat{\lambda}_{z_1}^{\mathrm{med}} = .75$, $\hat{\lambda}_{z_2}^{\mathrm{med}} = .69$), underscoring the tendency of cross-validation to remove such variables from the resulting estimate. Relative to the underspecified parametric model (Table 3), the proposed estimator exhibits a 14% predictive efficiency gain. The underspecification arises from neglected interaction terms, the omission of which is not uncommon in applied settings. Note also that the conventional frequency nonparametric estimator cannot remove irrelevant variables, and when there exist irrelevant variables, the efficiency loss of the conventional nonparametric method relative to the proposed cross-validated method is a substantial 12%.

Table 1. Correct Parametric Model: Median, 5th and 95th Percentiles of CCR

Conditional	*Frequency*	*Probit*
.815	.814	.853
[.744, .841]	[.757, .840]	[.821, .870]

5.2 Veterans Lung Cancer Data

We consider the dataset of Kalbfleisch and Prentice (1980, pp. 223–224) that 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 previous 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 the value 1 if survival is less than or equal to 180 days and 0 if 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 CCR for each split, and then summarize results over all 1,000 splits to alleviate any concerns that results from a single split may not be representative. Summarizing, 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 4 presents the median bandwidths over the 1,000 splits along with the 5th and 95th percentiles (with upper bounds given in square brackets).

Table 2. Overspecified Parametric Model: Median, 5th and 95th Percentiles of CCR

Conditional	*Frequency*	*Probit*
.652	.584	.629
[.564, .690]	[.527, .616]	[.576, .665]

Table 3. Underspecified Parametric Model: Median, 5th and 95th Percentiles of CCR

Conditional	Frequency	Probit
.815	.813	.715
[.748, .843]	[.754, .842]	[.689, .734]

It can be seen from Table 4 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. 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 as a prognosticator; patients with the highest (best) Karnofsky scores at the time of tumor diagnosis have the best survival and quality of life over the course of their illness.

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.

5.3 Female Labor Force Participation

The Mroz data file is taken from the 1976 panel study of income dynamics and is based on data for the previous year, 1975. Of the 753 observations, the first 428 are for women with positive hours worked in 1975, and the remaining 325 observations are for women who did not work for pay in 1975. (A more complete discussion of the data is given in Mroz 1987, app. 1.) The dataset consists of the following variables: LFP (a dummy variable that equals 1 if the woman worked in 1975 and 0 otherwise), KL6 (the number of children younger than age 6 years in the household), K618 (the number of children between ages 6 and 18), WA (wife's age), WE (wife's educational attainment, in years), CIT [a dummy variable that equals 1 if the family lives in a large city (SMSA) and 0 otherwise], UN (unemployment rate in county of residence, in percentage points and taken from bracketed ranges), LWW1 (logarithm of wife's average hourly earnings, in 1975 dollars for working women, log of predicted wage for nonworkers), PRIN (wife's property income computed as total family income minus the labor income earned by the wife).

We apply the proposed method for the prediction of labor force participation. We predict $\widehat{\text{LFP}} = 1$ if the estimated conditional probability that $\text{LFP} = 1$ exceeds .5, and $\widehat{\text{LFP}} = 0$ otherwise, and compare this with predictions from a logit model. For the full sample of size $n = 753$, we cross-validate and fit the conditional and logit densities, then predict the probability of labor force participation. To avoid the overfitting critique, we omit the ith individual's data point when predicting its probability of participation; that is, we form leave-one-out predictions for the kernel method. The confusion matrices (whose diagonal elements are correctly predicted outcomes and whose off-diagonal elements are incorrectly predicted outcomes) are presented in Table 5. It can be seen that the proposed method yields a 15% efficiency gain relative to the parametric logit specification, which is often used to model this dataset.

Table 5. Confusion Matrices

Logit			Conditional		
A/P	0	1	A/P	0	1
0	166	159	0	281	44
1	80	348	1	120	308
	CCR 68.2%			CCR 78.2%	

NOTE: A/P, actual/predicted values.

Next we consider the cross-validated bandwidths, and in particular focus on those for the categorical variables. The r in column 2 of Table 6 represents the number of values assumed by the associated categorical variables. Note from this table that the variables CIT and K618 were effectively "smoothed out" of the resulting estimate by the conditional density estimator, because their cross-validated bandwidths are fairly close to their maximum (upper-bound) values. It appears that the conditional cross-validation approach to bandwidth selection may provide substantial predictive improvements over the results obtained were one to use a logit model.

6. PROOF OF THEOREM 3

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 $C > 0$,

$$P\left(\sup\left|\frac{\hat{f}(\mathbf{x}, y)w(\mathbf{x}^{\mathrm{c}})}{\widehat{m}(\mathbf{x})} - \left[\frac{\mu_f(\mathbf{x}, y)w(\mathbf{x}^{\mathrm{c}})}{\mu_m(\mathbf{x})}\left\{1 - \frac{\Delta_m(\mathbf{x})}{\mu_m(\mathbf{x})}\right\} + \frac{\Delta_f(\mathbf{x}, y)w(\mathbf{x}^{\mathrm{c}})}{\mu_m(\mathbf{x})}\right]\right| > n^{-\delta}(nH)^{-1}\right) = O(n^{-C}), \tag{30}$$

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 $(\mathbf{x}, y)$ such that $\mathbf{x} = (\mathbf{x}^{\mathrm{c}}, \mathbf{x}^{\mathrm{d}})$, with $\mathbf{x}^{\mathrm{d}}$ in the support of the distribution of $\mathbf{X}^{\mathrm{d}}$; $h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q$ 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 $\epsilon \in (0, \frac{1}{2})$.

Table 4. Median Cross-Validated Bandwidth Values Over the 1,000 Splits, With Their 5th and 95th Percentiles in Parentheses

$\hat{\lambda}_1^{med}[.50]$	$\hat{\lambda}_2^{med}[.75]$	$\hat{\lambda}_3^{med}[.92]$	$\hat{\lambda}_4^{med}[.96]$	$\hat{\lambda}_5^{med}[.98]$	$\hat{\lambda}_6^{med}[.50]$
.50	.28	.01	.87	.96	.50
(.50, .50)	(.18, .36)	(0, .14)	(.76, .92)	(.72, .97)	(.20, .50)

NOTE: Numbers in square brackets represent maximum bandwidths, and so are the respective values of $(r_j - 1)/r_j$.

Table 6. Cross-Validated Bandwidths for Categorical Variables

Variable	r	Maximum	Conditional CV
CIT	2	.500	.492
KL6	4	.750	.095
K618	9	.888	.867
WA	31	.967	.822
WE	13	.923	0

Let η_n denote any positive sequence decreasing to 0, and let $\mathcal{S}_n$ be the set of all vectors, $\mathbf{v} = (h, h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$, 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 (30) implies that

$$\hat{g}(y|\mathbf{x}) = \frac{\mu_f(\mathbf{x}, y)}{\mu_m(\mathbf{x})}\left\{1 - \frac{\Delta_m(\mathbf{x})}{\mu_m(\mathbf{x})}\right\} - \frac{\Delta_f(\mathbf{x}, y)}{\mu_m(\mathbf{x})} + o_p\bigl(n^{-2/(p_1+5)}\bigr), \quad (31)$$

uniformly in $\mathbf{v} \in \mathcal{S}_n$. Here let $\mathbf{v} = \mathbf{v}_n$ denote any deterministic smoothing parameter sequence in $\mathcal{S}_n$. Property (28) is a consequence of (31) and the results

$$\frac{\mu_f(\mathbf{x}, y)}{\mu_m(\mathbf{x})} = \frac{\bar{\mu}_f(\bar{\mathbf{x}}, y)}{\bar{\mu}_m(\bar{\mathbf{x}})} = \bar{g}(y|\bar{\mathbf{x}}) + n^{-2/(p_1+5)}\beta(\bar{\mathbf{x}}, y) + o\bigl(n^{-2/(p_1+5)}\bigr), \quad (32)$$

$$\frac{n^{2/5}}{\nu_1(\mathbf{x})\bar{m}(\bar{\mathbf{x}})}\{\bar{g}(\bar{\mathbf{x}}, y)\Delta_m(\mathbf{x}) + \Delta_f(\mathbf{x}, y)\} \to \mathrm{N}\{0, \sigma(\bar{\mathbf{x}}, y)^2\}, \quad (33)$$

and

$$\left|\frac{\Delta_m(\mathbf{x})}{\nu_1(\mathbf{x})}\right| + \left|\frac{\Delta_f(\mathbf{x}, y)}{\nu_1(\mathbf{x})}\right| = O_p\bigl(n^{-2/(p_1+5)}\bigr), \quad (34)$$

where the convergence in (33) is in distribution. Property (32) follows by elementary calculus, (33) follows by a central limit theorem for sums of independent random variables, and (34) follows by calculating the mean square of the left side and showing that it equals $O(n^{-4/(p_1+5)})$.

We apply the superscript "$*$" to a quantity to denote the value that it takes for any particular deterministic but otherwise arbitrary $\mathbf{v} \in \mathcal{S}_n$. To extend (28) to the case where the smoothing parameter sequence is stochastic, and in particular obtained by cross-validation, it suffices to show that

$$\sup_{\mathbf{v}\in\mathcal{S}_n}\left|\frac{\bar{\mu}_f(\bar{\mathbf{x}}, y)}{\bar{\mu}_m(\bar{\mathbf{x}})} - \frac{\bar{\mu}_f^*(\bar{\mathbf{x}}, y)}{\bar{\mu}_m^*(\bar{\mathbf{x}})}\right| = o\bigl(n^{-2/(p_1+5)}\bigr), \quad (35)$$

$$\sup_{\mathbf{v}\in\mathcal{S}_n}\left|\frac{\Delta_f^*(\mathbf{x}, y)}{\nu_1^*(\mathbf{x})} - \frac{\Delta_f(\mathbf{x}, y)}{\nu_1(\mathbf{x})}\right| = o_p\bigl(n^{-2/(p_1+5)}\bigr), \quad (36)$$

and

$$\sup_{\mathbf{v}\in\mathcal{S}_n}\left|\frac{\Delta_m^*(\mathbf{x})}{\nu_1^*(\mathbf{x})} - \frac{\Delta_m(\mathbf{x})}{\nu_1(\mathbf{x})}\right| = o_p\bigl(n^{-2/(p_1+5)}\bigr). \quad (37)$$

Result (35) follows by elementary calculus, so it suffices to derive (36) and (37). We confine our attention to (36).

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

$$A_i(\mathbf{x}, y) = \left\{\prod_{j=1}^{p_1}\frac{1}{h_j}K\left(\frac{x_j^{\mathrm{c}} - X_{ij}^{\mathrm{c}}}{h_j}\right)\right\} \times \left\{\prod_{j=1}^{q}\left(\frac{\lambda_j}{r_j - 1}\right)^{N_{ij}(\mathbf{x})}(1 - \lambda_j)^{1 - N_{ij}(\mathbf{x})}\right\} \times \frac{1}{h}L\left(\frac{y - Y_i}{h}\right),$$

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

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

Elementary moment calculations show that $E(\Delta_f^\#)^2 = o(n^{-4/(p_1+5)})$, uniformly in smoothing parameters $\mathbf{v} \in \mathcal{S}_n$. Using properties of rates of convergence in invariance principles for multivariate empirical processes (see, e.g., Rio 1996), these results may be generalized by showing that the normalized stochastic process, indexed by $v \in \mathcal{S}$, converges to 0 uniformly in smoothing parameters in $\mathcal{S}_n$,

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

in probability.

Therefore, to establish (36) it suffices to prove that

$$\sup_{\mathbf{v}\in\mathcal{S}_n}\bigl|\Delta_f^\dagger(\mathbf{x}, y) - \Delta_f^{\dagger *}(\mathbf{x}, y)\bigr| = o\bigl(n^{-2/(p_1+5)}\bigr), \quad (38)$$

where $\Delta_f^{\dagger *}(\mathbf{x}, y)$ denotes the version of $\Delta_f^\dagger(\mathbf{x}, y)$ computed for any particular value of the smoothing parameter vector $\mathbf{v}$. [Note that neither $\Delta_f^{\dagger *}(\mathbf{x}, y)$ nor $\Delta_f^\dagger(\mathbf{x}, y)$ depends on the last p_2 h_j's or the last q_2 λ_j's.] Simple moment calculations show that $E\{\Delta_f^\dagger(\mathbf{x}, y) - \Delta_f^{\dagger *}(\mathbf{x}, y)\}^2 = o(n^{-4/(p_1+5)})$ uniformly in $\mathbf{v} \in \mathcal{S}_n$, and this result again may be extended, to (38), using properties of invariance principles for multivariate empirical processes. Therefore (36) holds, completing the proof of the theorem.

[Received July 2003. Revised April 2004.]

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. (2004), "A Data-Driven Method for Estimating Conditional Densities," Research Report 2003-05, Institute of Mathematical Sciences, Chinese University of Hong Kong.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984), "Projection Pursuit Density Estimation," *Journal of the American Statistical Association*, 79, 599–608.

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, Vienna: Physica-Verlag, pp. 101–110.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hall, P. (1983a), "Orthogonal Series Methods for Both Qualitative and Quantitative Data," *The Annals of Statistics*, 11, 1004–1007.

——— (1983b), "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics* 11, 1156–1174.

——— (1985), "Asymptotic Theory of Minimum Integrated Square Error for Multivariate Density Estimation," in *Multivariate Analysis VI*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 289–309.

Hall, P., and Marron, J. S. (1987), "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," *Probability Theory and Related Fields*, 74, 567–581.

Hall, P., and Titterington, D. M. (1987), "On Smoothing Sparse Multinomial Data," *Australian Journal Statistics*, 29, 19–37.

Hirano, K., Imbens, G., and Ridder, G. (2002), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horowitz, J. (2001), "Nonparametric Estimation of a Generalized Additive Model With an Unknown Link Function," *Econometrica*, 69, 499–513.

Huber, P. J. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435–525.

Jones, M. C., and Sibson, R. (1987), "What Is Projection Pursuit?" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 150, 1–36.

Kalbfleisch, J., and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

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, 765–779.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Li, Q., and Racine, J. S. (2003), "Nonparametric Estimation of Joint Distribution With Mixed Continuous and Categorical Data," *Journal of Multivariate Analysis*, 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," *Comptes Rendus de l'Academie des Sciences*, *Série I Math.*, 322, 169–172.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, 9, 65–78.

Sain, S. R. (2001), "Bias Reduction and Elimination With Kernel Estimators," *Communications in Statistics Theory Methods*, 30, 1869–1888.

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

Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, 12, 1285–1297.

Tutz, G. (1991), "Sequential Models in Categorical Regression," *Computational Statistics and Data Analysis*, 11, 275–295.

Wang, M. C., and van Ryzin, J. A. (1981), "A Class of Smooth Estimators for Discrete Distributions," *Biometrika*, 68, 301–309.

This article has been cited by:

1. Sam Efromovich. 2010. Dimension Reduction and Adaptation in Conditional Density EstimationDimension Reduction and Adaptation in Conditional Density Estimation. *Journal of the American Statistical Association* **105**:490, 761-774. [Abstract] [PDF] [PDF Plus]
2. Song Xi Chen, Cheng Yong Tang, Vincent T. Mule, Jr.. 2010. Local Post-Stratification in Dual System Accuracy and Coverage Evaluation for the U.S. CensusLocal Post-Stratification in Dual System Accuracy and Coverage Evaluation for the U.S. Census. *Journal of the American Statistical Association* **105**:489, 105-119. [Abstract] [PDF] [PDF Plus]
3. Yacine Aït-Sahalia, Jianqing Fan, Heng Peng. 2009. Nonparametric Transition-Based Tests for Jump DiffusionsNonparametric Transition-Based Tests for Jump Diffusions. *Journal of the American Statistical Association* **104**:487, 1102-1116. [Abstract] [PDF] [PDF Plus] [Supplementary material]
4. Qi Li , Jeffrey S. Racine , Jeffrey M. Wooldridge . 2009. Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous DataEfficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data. *Journal of Business and Economic Statistics* **27**:2, 206-223. [Abstract] [PDF] [PDF Plus]
5. Qi Li , Jeffrey S . Racine . 2008. Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous DataNonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data. *Journal of Business and Economic Statistics* **26**:4, 423-434. [Abstract] [PDF] [PDF Plus]