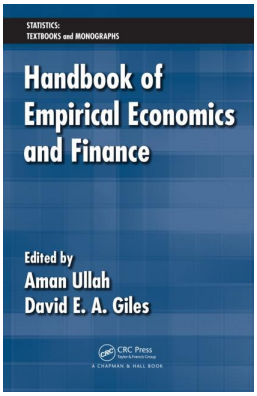

Handbook of Empirical Economics and Finance

Ullah Aman, E. A. Giles David

Nonparametric Kernel Methods for Qualitative and Quantitative Data

Publication details
https://test.routledgehandbooks.com/doi/10.1201/b10440-8
S. Racine Jeffrey
Published online on: 20 Dec 2010

7

Nonparametric Kernel Methods for Qualitative and Quantitative Data

Jeffrey S. Racine

CONTENTS

7.1 Introduction

Nonparametric kernel methods have become an integral part of the applied econometrician's toolkit. Their appeal, for applied researchers at least, lies in their ability to reveal structure in data that might be missed by classical parametric methods. Basic kernel methods are now found in virtually all

popular statistical and econometric software programs. Such programs contain routines for the estimation of an unknown density function defined over a real-valued continuous random variable, or for the estimation of an unknown bivariate regression model defined over a real-valued continuous regressor. For example, the R platform for statistical computing and graphics (R Development Core Team 2008) includes the function `density` that computes a univariate kernel density estimate supporting a variety of kernel functions and bandwidth methods, while the `locpoly` function in the R "KernSmooth" package (Wand and Ripley 2008) can be used to estimate a bivariate regression function and its derivatives using a local polynomial kernel estimator with a fast binned bandwidth selector.

Those familiar with traditional nonparametric kernel smoothing methods such as that embodied in `density` or `locpoly` will appreciate that these methods presume that the underlying data are real-valued and continuous in nature, which is frequently not the case as one often encounters categorical along with continuous data types in applied settings. A popular traditional method for handling the presence of both continuous and categorical data is called the "frequency" approach. For this approach the data are first broken up into subsets ("cells") corresponding to the values assumed by the categorical variables, and then one applies, say, `density` or `locpoly` to the continuous data remaining in each cell. Unfortunately, nonparametric frequency approaches are widely acknowledged to be unsatisfactory because they often lead to substantial efficiency losses arising from the use of sample splitting, particularly when the number of cells is large.

Recent developments in kernel smoothing offer applied econometricians a range of kernel-based methods for categorical data only (i.e., unordered and ordered factors), or for a mix of continuous and categorical data. These methods have the potential to recover the efficiency losses associated with nonparametric frequency approaches since they do not rely on sample splitting. Instead, they smooth the categorical variables in an appropriate manner; see Li and Racine (2007) and the references therein for an in-depth treatment of these methods, and see also the references listed in the bibliography.

In this chapter we shall consider a range of kernel methods appropriate for the mix of categorical and continuous data one often encounters in applied settings. Though implementations of hybrid methods that admit the mix of categorical and continuous data types are quite limited, there exists an R package titled "np" (Hayfield and Racine 2008) that implements a variety of hybrid kernel methods, and we shall use this package to illustrate a few of the methods that are discussed in the following sections. Since many readers will no doubt be familiar with the classical approaches embodied in the functions `density` or `locpoly` or their peers, we shall begin with some recent developments in the kernel smoothing of categorical data only.

7.2 Kernel Smoothing of Categorical Data

The kernel smoothing of categorical data would appear to date from the seminal work of Aitchison and Aitken (1976) who proposed a novel method for kernel estimation of a probability function defined over multivariate binary data types. The wonderful monograph by Simonoff (1996) also contains chapters on the kernel smoothing of categorical data types such as sparse contingency tables and so forth. Econometricians are more likely than not interested in estimation of conditional objects, so we shall introduce the kernel smoothing of categorical objects via the estimation of a probability function and then immediately proceed to the estimation of a conditional mean. The estimation of a conditional mean with categorical covariates offers a unique springboard for presenting recent developments that link kernel smoothing to Bayesian methods. This exciting development offers a deeper understanding of kernel methods while also delivering novel methods for bandwidth selection and provides bounds ensuring that kernel smoothing will dominate frequency methods on mean square error (MSE) grounds.

7.2.1 Kernel Smoothing of Univariate Categorical Probabilities

Suppose we were interested in estimating a univariate *probability* function where the data are categorical in nature. The nonparametric nonsmooth approach would construct a frequency estimate, while the nonparametric smooth approach would construct a kernel estimate. For those unfamiliar with the term "frequency" estimate, we mean simply the estimator of a probability computed via the sample frequency of occurrence. For example, if a random variable is the result of a Bernoulli trial (i.e., zero or one with fixed probability from trial to trial) then the frequency estimate of the probability of a zero (one) is simply the number of zeros (ones) divided by the number of trials.

First, consider the estimation of a probability function defined for $X_i \in \mathcal{S} = \{0, 1, \ldots, c-1\}$. The nonsmooth "frequency" (nonkernel) estimator of $p(x)$ is given by

$$\tilde{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i, x),$$

where $\mathbf{1}(A)$ is an indicator function taking on the value 1 if A is true, zero otherwise. It is straightforward to show that

$$E\,\tilde{p}(x) = p(x),$$
$$\text{Var}\ \tilde{p}(x) = \frac{p(x)(1-p(x))}{n},$$

hence,

$$\text{MSE}(\tilde{p}(x)) = n^{-1}p(x)(1-p(x)) = O(n^{-1}),$$

which implies that

$$\tilde{p}(x) - p(x) = O_p\left(n^{-1/2}\right).$$

Now, consider the kernel estimator of $p(x)$,

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, x, \lambda), \tag{7.1}$$

where $l(\cdot)$ is a kernel function defined by, say,

$$l(X_i, x, \lambda) = \begin{cases} 1-\lambda & \text{if } X_i = x \\ \lambda/(c-1) & \text{otherwise,} \end{cases} \tag{7.2}$$

and where $\lambda \in [0, (c-1)/c]$ is a "smoothing parameter" or "bandwidth." The requirement that λ lie in $[0, (c-1)/c]$ ensures that $\tilde{p}(x)$ is a proper probability estimate lying in $[0, 1]$. It is easy to show that

$$\begin{aligned} E\,\hat{p}(x) &= p(x) + \lambda\left\{\frac{1-cp(x)}{c-1}\right\}, \\ \text{Var}\hat{p}(x) &= \frac{p(x)(1-p(x))}{n}\left(1-\lambda\frac{c}{(c-1)}\right)^2. \end{aligned} \tag{7.3}$$

This estimator was proposed by Aitchison and Aitken (1976) for discriminant analysis with multivariate binary data; see also Simonoff (1996).

The above expressions indicate that the kernel smoothed estimator may possess some finite-sample bias; however, its finite-sample variance is less than its frequency counterpart. This suggests that the kernel estimator can dominate the frequency estimator on MSE grounds, which turns out to be the case; see Ouyang, Li, and Racine (2006) for extensive simulations. Results similar to those outlined in Subsection 7.3.1 for categorical Bayesian methods could be extended to this setting, though we do not attempt this here for the sake of brevity.

Note that when $\lambda = 0$ this estimator collapses to the frequency estimator $\tilde{p}(x)$, while when λ hits its upper bound, $(c-1)/c$, this estimator is the rectangular (i.e., discrete uniform) estimator which yields equal probabilities across all outcomes.

Using a bandwidth that balances bias and variance such as that proposed by Ouyang, Li, and Racine (2006), it can be shown that

$$\hat{p}(x) - p(x) = O_p\left(n^{-1/2}\right).$$

It can also be shown that

$$\sqrt{n}(\hat{p}(x) - p(x)) \to N\{0, p(x)(1 - p(x))\} \text{ in distribution.} \tag{7.4}$$

See Ouyang, Li, and Racine (2006) for details. For the sake of brevity we shall gloss over bandwidth selection methods, and direct the interested reader to Ouyang, Li, and Racine (2006) and Li and Racine (2007) for a detailed description of data-driven bandwidth selection methods for this object.

We have considered the univariate estimator by way of introduction. A multivariate version follows trivially by replacing the univariate kernel function with a multivariate product kernel function. We would let X now denote an r-dimensional discrete random vector taking values on $\mathcal{S}$, the support of X. We use x^s and X_i^s to denote the sth component of x and X_i $(i = 1, \ldots, n)$, respectively. The product kernel function is then given by

$$L_\lambda(X_i, x) = \prod_{s=1}^{r} l(X_i^s, x^s, \lambda_s) = \prod_{s=1}^{r} \{\lambda_s/(c_s - 1)\}^{I_{x_i^s \neq x^s}} (1 - \lambda_s)^{I_{x_i^s = x^s}}, \tag{7.5}$$

where $I_{x_i^s \neq x^s} = I(X_i^s \neq x^s)$, and $I_{x_i^s = x^s} = I(X_i^s = x^s)$. The kernel estimator is identical to that in Equation 7.1 except that we replace $l(X_i, x, \lambda)$ with $L_\lambda(X_i, x)$. All results (rate of convergence, asymptotic normality, etc.) remain unchanged.

7.2.1.1 A Simulated Example

In the following R code chunk we simulate $n = 250$ draws from five trials of a Bernoulli process having probability of success 1/2 from trial to trial, hence $x \in \{0, \ldots, 5\}$ and $c = 6$.

```
R> library("np")

Nonparametric Kernel Methods for Mixed Datatypes
   (version 0.30-7)

R> library(xtable)
R> set.seed(12345)
R> n <- 250
R> x <- sort(rbinom(n,5,.5))
R> ## Compute the non-smoothed (frequency) probability
   estimates
R> ptilde <- table(x)/n
R> ## Compute the smoothed probability estimates
R> phat <- unique(fitted(npudens(~factor(x))))
```

It can be seen that the nonsmooth frequency and the smooth kernel estimates are quite close for this example as expected, while the kernel estimators shrink slightly toward the uniform probability estimate

TABLE 7.1

Nonparametric Frequency ($\tilde{p}(x)$, Nonsmooth) and Nonparametric Smoothed ($\hat{p}(x)$) Probability Estimates.

x	$\tilde{p}(x)$	$\hat{p}(x)$
0	0.024	0.029
1	0.132	0.133
2	0.272	0.268
3	0.360	0.353
4	0.168	0.168
5	0.044	0.049

$p = 1/c = 1/6 = 0.1667$. We shall discuss the relationship between the kernel estimator and Bayesian methods in Subsection 7.3.1.

7.2.2 Kernel Smoothing of Bivariate Categorical Conditional Means

Now suppose by way of example that we observe $\{Y_i, X_i\}$ pairs generated by $y = g(x) + \epsilon$, where $g(x)$ is defined by

$$Y_i = X_i + \epsilon_i \tag{7.6}$$

where $X_i \in \mathcal{S} = \{0, 1, \ldots, c-1\}$ and $\epsilon_i \sim N(0, 1)$ represent i.i.d. draws.

The nonsmooth "frequency" (nonkernel) estimator of $g(x)$ (which is also the least squares estimator) is given by

$$\tilde{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i, x)}{\sum_{i=1}^{n} \mathbf{1}(X_i, x)},$$

which simply returns the sample mean of those Y_i for which $X_i = x \in \mathcal{S} = \{0, 1, \ldots, c-1\}$. It can be shown that

$$\tilde{g}(x) - g(x) = O_p\left(n^{-1/2}\right).$$

Now, consider the kernel estimator of $g(x)$,

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i l(X_i, x, \lambda)}{\sum_{i=1}^{n} l(X_i, x, \lambda)}, \tag{7.7}$$

where $l(\cdot)$ is, say, the kernel function defined in Equation 7.2.

Note that when $\lambda = 0$ this estimator collapses to the frequency estimator $\tilde{g}(x)$, while when λ hits its upper bound, $(c-1)/c$, this estimator yields equal fitted values across all $x \in \mathcal{S} = \{0, 1, \ldots, c-1\}$, namely, the overall (unconditional) mean of Y_i.

Using a bandwidth that balances bias and variance, it can be shown that

$$\hat{g}(x) - g(x) = O_p\left(n^{-1/2}\right),$$

TABLE 7.2

Nonparametric Frequency ($\tilde{g}(x)$, Nonsmooth) and Nonparametric Smoothed ($\hat{g}(x)$) Regression Estimates

x	$\tilde{g}(x)$	$\hat{g}(x)$
0	−0.587	−0.484
1	0.860	0.871
2	2.092	2.094
3	3.055	3.054
4	4.072	4.066
5	5.574	5.524

and that

$$\sqrt{n}\left(\hat{g}(x) - g(x)\right) / \sqrt{\hat{\Omega}(x)} \to N(0, 1) \text{ in distribution,}$$

where $\hat{\Omega}(x) = \hat{\sigma}^2(x)/\hat{p}(x)$, and where $\hat{\sigma}^2(x) = n^{-1}\sum_i [Y_i - \hat{g}(X_i)]^2 l(X_i, x, \hat{\lambda})/\hat{p}(x)$ is a consistent estimator of $\sigma^2(x) = E(u_i^2 \mid X_i = x)$. See Ouyang, Li, and Racine (2008) for details.

7.2.2.1 A Simulated Example

In the following R code chunk we simulate $n = 250$ draws for x from five trials of a Bernoulli process having probability of success 1/2 from trial to trial, hence $x \in \{0, \ldots, 5\}$ and $c = 6$, then simulate $y = x + \epsilon$ where $\epsilon \sim N(0, 1)$.

```
R> set.seed(12345)
R> n <- 250
R> x <- sort(rbinom(n,5,.5))
R> y <- x + rnorm(n)
R> ## Regression on dummy variables (same as unconditio-
   nal group means)
R> gtilde <- unique(predict(model.par <- lm(y~factor(x))
   ))
R> ## Nonparametric regression on a factor (shrink
   towards overall mean)
R> ghat <- unique(predict(model.np <- npreg(y~factor(x))
   ))
```

We have considered the univariate estimator by way of introduction. A multivariate version follows trivially by replacing the univariate kernel function with a multivariate product kernel function defined in Equation 7.5. The kernel estimator is identical to that in Equation 7.7 except that we replace $l(X_i, x, \lambda)$ with $L_\lambda(X_i, x)$. All results (rate of convergence, asymptotic normality, etc.) remain unchanged.

7.3 Categorical Kernel Methods and Bayes Estimators

Kiefer and Racine (2009) have recently investigated the relationship between nonparametric categorical kernel methods and hierarchical Bayes models of the type considered by Lindley and Smith (1972). By exploiting certain similarities among the approaches, they gain a deeper understanding of the nature of kernel-based methods and leverage some theoretical apparatus developed for hierarchical Bayes models which is immediately relevant for kernel-based techniques. We outline their approach below as it provides additional insight and also delivers a new approach toward bandwidth selection for categorical kernel methods.

7.3.1 Kiefer and Racine's (2009) Analysis

In order to facilitate a direct comparison with Kiefer and Racine's (2009) notation, we now let the sample realizations $\{X_i, Y_i\}$ be written instead as $\{X_{ji}, Y_{ji}\}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, c$. We let y_i be the frequency estimator of μ_i defined as

$$y_i = \frac{1}{n_i} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} \mathbf{1}(X_{jk} = i), \tag{7.8}$$

i.e., the sample mean of Y when $X = i$ (a "cell" mean). Let $y_{\bar{i}}$ be defined as

$$y_{\bar{i}} = \frac{1}{(n - n_i)} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} \mathbf{1}(X_{jk} \neq i),$$

i.e., the sample mean of Y over all values of X other than $X = i$ ($\bar{i}$ is taken to be the complement of i), while the frequency estimator of $E(Y)$ (the "overall" mean) is

$$y_{\cdot} = \frac{1}{n} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} = \frac{n_i y_i + (n - n_i) y_{\bar{i}}}{n}.$$

Adopting Kiefer and Racine's (2009) notation, the kernel estimator of μ_i could be written as

$$y_{i,\lambda} = \hat{g}(i) = \frac{n^{-1} \sum_{k=1}^{c} \sum_{j=1}^{n_k} Y_{jk} L(X_{jk}, i, \lambda)}{p_{i,\lambda}}.$$

In order to facilitate a comparison of the Bayesian approach of Lindley and Smith (1972) and the kernel approach, we wish to express $y_{i,\lambda}$ as a weighted

average of y_i and $y_.$. The kernel estimator $y_{i,\lambda}$ can be rewritten as follows,

$$\begin{aligned} y_{i,\lambda} &= \frac{n^{-1}\sum_{k=1}^{c}\sum_{j=1}^{n_k} Y_{jk} L(X_{jk}, i, \lambda)}{p_{i,\lambda}} \\ &= \frac{n^{-1}\left(n_i y_i (1-\lambda) + (n-n_i) y_{\bar{i}} \lambda/(c-1)\right)}{n^{-1}\left(n_i(1-\lambda) + (n-n_i)\lambda/(c-1)\right)} \\ &= \frac{n_i y_i (1-\lambda) + (n y_. - n_i y_i)\,\lambda/(c-1)}{n_i(1-\lambda) + (n-n_i)\lambda/(c-1)} \\ &= \left[\frac{n_i/n\,(1-\lambda c/(c-1))}{n_i/n\,(1-\lambda c/(c-1)) + \lambda/(c-1)}\right] y_i \\ &\quad + \left[\frac{\lambda/(c-1)}{n_i/n\,(1-\lambda c/(c-1)) + \lambda/(c-1)}\right] y_. \\ &= (1-\Phi_i) y_i + \Phi_i y_., \end{aligned}$$

where the third equality follows from Equation 7.8 by noting that

$$n y_. - n_i y_i = (n-n_i) y_{\bar{i}},$$

where

$$1-\Phi_i = \left[\frac{n_i/n\,(1-\lambda c/(c-1))}{n_i/n\,(1-\lambda c/(c-1)) + \lambda/(c-1)}\right]$$

and

$$\Phi_i = \left[\frac{\lambda/(c-1)}{n_i/n\,(1-\lambda c/(c-1)) + \lambda/(c-1)}\right],$$

and where $\lambda \in [0, (c-1)/c]$ implies that $\Phi_i \in [0, 1]$.

When $\lambda = 0$ (i.e., $\Phi_i = 0 \forall i$), $y_{i,\lambda} = y_i$ (the frequency estimator), while when $\lambda = (c-1)/c$ (i.e., $(1-\lambda c/(c-1)) = 0$ or $\Phi_i = 1 \forall i$), $y_{i,\lambda} = y_., i = 1, \ldots, c$ (the global mean). Note that this is exactly the same result using the notation in Equation 7.7.

Kiefer and Racine (2009) consider hierarchical models of the form

$$y_{ji} = \mu_i + \epsilon_{ji}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, c,$$

where n_i is the number of observations drawn from group i, and where there exist c groups.

For the ith group,

$$\begin{pmatrix} y_{1i} \\ \vdots \\ y_{n_i i} \end{pmatrix} = \iota_{n_i} \mu_i + \epsilon_i, \quad i = 1, \ldots, c,$$

where ι_{n_i} is a vector of ones of length n_i, $\epsilon_i = (\epsilon_{1i}, \ldots, \epsilon_{n_i i})'$, and, for the sample, $\mathbf{y} = A\mu + \epsilon$ where $\mathbf{y}$ is the n-vector of observations, A is the $(n \times c)$ design matrix, and $\mu = (\mu_1, \ldots, \mu_c)'$, the vector of group means. The goal is to understand the connection between hierarchical Bayes models and kernel estimators of multivariate means.

Kiefer and Racine (2009) consider a three-stage hierarchical Bayes model. The first stage is given by

$$\mathbf{y} \sim (A_1\theta_1, C_1).$$

As a function of θ_1 and C_1 for given y, this first stage specification can be regarded as the likelihood function for the normally distributed case, otherwise as a quasi likelihood based on two moments (Heyde 1997). We return to A_1 below.

The second stage,

$$\theta_1 \sim (A_2\theta_2, C_2),$$

can be regarded as a prior distribution for θ_1 given $A_2\theta_2$ and C_2 in the normal case (where it is conjugate) or as an approximation to the prior if not normal, or from a frequency viewpoint as a second stage in the data generating process (DGP). The first stage "parameters" are themselves generated by a random process in this view. This interpretation focuses attention on the hyperparameters θ_2 (and C_2) rather than θ_1, which strictly speaking is not a parameter in the frequency sense.

The third stage,

$$\theta_2 \sim (A_3\theta_3, C_3),$$

can again be regarded as a prior on the second stage parameter θ_2, or as an additional stage in the DGP.

Interest lies in estimating the $c \times 1$ vector of means θ_1. Following Lindley and Smith (1972) we are thinking of normal distributions at each stage. For our purposes we can also regard the stages as approximate distributions characterized by two moments noting the calculations are exact only for the normal. The point of the stages is that the dimension of the conditioning parameter is reduced at each step. We are using the Bayesian hierarchical setup to obtain insight into the kernel estimator. Lindley and Smith (1972) suggest specifications proportional to identity matrices and inverted gamma densities for the factors of proportion (and related generalizations). They suggest using modal estimators in the expressions for the posterior means of interest.

For the problem at hand, we try to stick with the notation of Lindley and Smith as closely as possible. The first stage is

$$A_1 = \{a_{ji}\} \text{ with } a_{ji} \in \{0, 1\}, \sum_{i=1}^{c} a_{ki} = 1, \sum_{k=1}^{n} a_{ki} = n_i,$$

$$\theta_1 = \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_c \end{bmatrix},$$

$$C_1 = \sigma^2 I_n.$$

A_1 is the $n \times c$ design matrix with $A_1'A_1$ the $c \times c$ diagonal matrix with n_i, the number of observations in the ith group, as the ith diagonal element, μ is a $c \times 1$ vector of (population) group means, σ^2 is the within-group variance (i.e., $\text{Var}(y_{ij})$), and I_n is the $n \times n$ identity matrix. Next, the second stage will become

$$A_2 = \iota_c,$$

$$\theta_2 = \mu_.,$$

$$C_2 = \tau^2 I_c,$$

where $\mu_.$ is the (population) "overall mean," and $\tau^2 = \text{Var}(\mu_i)$. Note that $A_2\theta_2 = \iota\mu_.$ is simply a $c \times 1$ vector with elements being the overall mean $\mu_.$ to which the Bayes (and kernel) estimators can shrink. Finally, we let the scalar

$$C_3^{-1} \to 0$$

so that the prior on $\mu_.$ is improper. Note that the impropriety is confined to one dimension. The frequency analysis corresponds to an improper prior on the c-vector θ_1, so that we expect inadmissibility of the frequency estimator through a Stein effect if $c > 2$. By adding a third stage, we reduce the improper prior to one dimension. The results are seen below.

The three stage Bayes estimate is (Lindley and Smith 1972, p. 7, Eq. 16)

$$\theta_1^* = D_0 d_0$$

where

$$D_0^{-1} = \left(A_1'C_1^{-1}A_1 + C_2^{-1} - C_2^{-1}A_2\left(A_2'C_2^{-1}A_2\right)^{-1}A_2'C_2^{-1}\right)$$

$$d_0 = \left(A_1'C_1^{-1}y\right).$$

θ_1^* is the posterior mean and is an optimal estimator under quadratic loss. Writing

$$\Lambda = A_1' C_1^{-1} A_1 = \frac{1}{\sigma^2} \begin{bmatrix} n_1 & 0 & 0 & \dots \\ 0 & n_2 & 0 & \dots \\ \vdots & & \ddots & \\ \vdots & 0 & 0 & n_c \end{bmatrix},$$

we see that

$$\begin{aligned} D_0^{-1} &= \left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota (\iota'^{-2} \iota)^{-1} \iota'^{-2}\right) \\ &= \left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota\iota'/c\right), \\ d_0 &= A_1' C_1^{-1} \mathbf{y} \\ &= \begin{pmatrix} \frac{y_1 n_1}{\sigma^2} \\ \vdots \\ \frac{y_c n_c}{\sigma^2} \end{pmatrix}. \end{aligned}$$

Recall that y_i is the mean for group i. Thus the vector of posterior means satisfies

$$\left(\Lambda + \tau^{-2} I_c - \tau^{-2} \iota (\iota'\iota)^{-1} \iota'\right) \theta_1^* = d_0$$

or, element-wise

$$(\sigma^{-2} n_j + \tau^{-2}) \theta_{1j}^* - \tau^{-2} \theta_{1.}^* = \sigma^{-2} n_j y_j,$$

where $\theta_{1.}^* = \sum_{j=1}^{c} \theta_{1j}^*/c$. Thus

$$\theta_{1j}^* = (\sigma^{-2} n_j y_j + \tau^{-2} \theta_{1.}^*)/(\sigma^{-2} n_j + \tau^{-2})$$

and the Bayes estimator for the jth mean is a weighted average of the group mean and the overall posterior mean. This, in general, cannot be expressed as a weighted average of the group mean and the overall mean.

We consider the "balanced case" (n_i equal for all i) in what follows. Let $n_i = n^*$ for all i. The kernel estimator of the ith component of μ can be

written as

$$\begin{aligned} y_{i,\lambda} &= \left[\frac{n^*(1-\lambda c/(c-1))}{n^*(1-\lambda c/(c-1))+n\lambda/(c-1)}\right] y_i \\ &\quad + \left[\frac{n^*\lambda/(c-1)}{n^*(1-\lambda c/(c-1))+n^*\lambda/(c-1)}\right] y_{\cdot} \\ &= \left[\frac{n^*}{n^*+n^*/((c-1)/\lambda-c)}\right] y_i + \left[\frac{n^*/((c-1)/\lambda-c)}{n^*+n^*/((c-1)/\lambda-c)}\right] y_{\cdot}, \end{aligned} \tag{7.9}$$

where λ is a smoothing parameter to be set by the researcher.

Further, the Bayes estimator of the ith component of μ is given by (in the balanced case)

$$\begin{aligned} \mu_i^* &= \left[\frac{n^*}{n^*+\kappa^{-1}}\right] y_i + \left[\frac{\kappa^{-1}}{n^*+\kappa^{-1}}\right] y_{\cdot} \\ &= v y_i + (1-v) y \end{aligned} \tag{7.10}$$

where $v = n^*/(n^*+\kappa^{-1})$ is the common value of the v_i term from above. The correspondence between the two methods is given by

$$n^*/((c-1)/\lambda - c) = \kappa^{-1},$$

hence

$$\kappa = \frac{1}{n^*}((c-1)/\lambda - c).$$

Alternatively, λ can be expressed as

$$\lambda = (c-1)/(c+n^*\kappa). \tag{7.11}$$

This gives some intuition for the choice of the smoothing parameter λ if one chooses not to adopt the Bayesian approach explicitly. λ should be larger as the groups are thought to be more homogeneous (smaller κ or τ^2) and smaller as the groups are thought to be less similar.

Next, we turn to another frequency property, that of MSE. It is known that the MSE of the Bayes/kernel estimator (identical in the balanced case) improves over that of the frequency estimator y_i if and only if (Lindley and Smith 1972, p. 3, Eq. 2)

$$\hat{\tau}^2 \leq 2\tau^2 + \sigma^2,$$

where

$$\hat{\tau}^2 = \sum_i \frac{(y_i - y_{\cdot})^2}{c-1}. \tag{7.12}$$

This allows us to obtain an upper bound for λ that will ensure (in probability) that $\text{MSE}(y_{i,\lambda}) \leq \text{MSE}(y_i)$. Substituting, we have

$$\hat{\tau}^2 \leq 2\frac{\sigma^2}{n}((c-1)/\lambda - c) + \sigma^2,$$

which is equivalent to

$$\frac{n(\hat{\tau}^2 - \sigma^2)}{2\sigma^2} + c \leq \frac{c-1}{\lambda},$$

which implies that

$$\lambda \leq \frac{2\sigma^2(c-1)}{n(\hat{\tau}^2 - \sigma^2) + 2c\sigma^2}. \tag{7.13}$$

The only unknown in this formula is σ^2, which can be estimated directly from the data via

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n^*}(y_{ij} - y_i)^2}{n-c}. \tag{7.14}$$

It is widely known that the smoothing parameter must obey $\lambda \to 0$ as $n \to \infty$ for consistent estimation while, as noted earlier, λ is restricted to lie in $[0, (c-1)/c]$ (see Equation 7.2). Note that Equation 7.13 tells us that an *oversmoothed* kernel estimator can be consistent but can be beaten by the frequency estimator on MSE grounds (i.e., when λ is overly large).

The results obtained above yield a number of implications for applied kernel estimation with categorical data. The first is that they provide bounds for bandwidth selection that are previously unknown in the literature. The second is that they deliver a simple plug-in method of bandwidth selection with an empirical Bayes flavor (Efron and Morris 1973) that possesses appealing finite-sample properties and, in addition, is computationally trivial. Recall that $[0, (c-1)/c]$ is the range of λ when using the kernel function defined in Equation 7.2. We now incorporate the result summarized in Equation 7.13 to obtain tighter bounds on λ.

Note that when $\hat{\tau}^2 = \sigma^2$, Equation 7.13 equals $(c-1)/c$, the upper bound possible for λ, hence the bound is nonbinding in this case. It is also nonbinding when $\hat{\tau}^2 \leq \sigma^2$. However, when $\hat{\tau}^2 > \sigma^2$, then in order to outperform the frequency estimator on MSE grounds, the kernel estimator must obey $\lambda < (c-1)/c$ with the upper bound now given by Equation 7.13. On MSE grounds, the range of λ is no longer $[0, (c-1)/c]$, rather it is

$$\left[0, \min\left\{\frac{c-1}{c}, \frac{2\sigma^2(c-1)}{n(\hat{\tau}^2 - \sigma^2) + 2c\sigma^2}\right\}\right]. \tag{7.15}$$

In other words, Equation 7.13 tells us that when the idiosyncratic variation (i.e., $\sigma^2 = \text{Var}(y_{ij})$) is greater than the intergroup variation (i.e., $\hat{\tau}^2 = \text{Var}(y_i)$), there exists a λ in the feasible range (i.e., $[0, (c-1)/c]$) that will outperform the

frequency estimator on MSE grounds (e.g., that given by Equation 7.11). On the other hand, when the idiosyncratic variation is less than the intergroup variation, imposing this (reduced) bound on λ (rather than $(c-1)/c$) avoids situations where the frequency estimator may outperform the smoothed estimator. Note that Equation 7.11 always satisfies the bound.

Equation 7.11 suggests a computationally trivial formula for a plug-in bandwidth selector for the kernel estimator of a multivariate mean that might serve as an alternative to that proposed in Ouyang, Li, and Racine (2008).

7.3.1.1 A Simulated Example

Next we simulate $y = \epsilon$, where $\epsilon \sim N(0, 1)$, and use leave-one-out cross-validation to select the unknown bandwidth.

```
R> set.seed(12345)
R> n <- 250
R> x <- sort(rbinom(n,5,.5))
R> y <- rnorm(n)
R> ## Regression on dummy variables (same as unconditio-
   nal group means)
R> gtilde <- unique(predict(model.par <- lm(y~factor(x))
   ))
R> ## Nonparametric regression on a factor (shrink towa-
   rds overall mean)
R> ghat <- unique(predict(model.np <- npreg(y~factor(x))
   ))
```

Note that, for this example, the unconditional mean of y is $y_{\cdot} = 0.05$. It can be seen from the above example that the kernel estimator correctly shrinks the nonparametric frequency estimator towards the overall mean in accordance with the findings of Kiefer and Racine (2009).

We now discuss recent developments in the kernel estimation of objects involving the mix of categorical and continuous data types often found in applied settings.

TABLE 7.3

Nonparametric Frequency ($\tilde{g}(x)$, Nonsmooth) and Nonparametric Smoothed ($\hat{g}(x)$) Regression Estimates.

x	$\tilde{g}(x)$	$\hat{g}(x)$
0	−0.587	0.050
1	−0.140	0.050
2	0.092	0.050
3	0.055	0.050
4	0.072	0.050
5	0.574	0.050

7.4 Kernel Methods with Mixed Data Types

So far we have presumed that the categorical variable is of the "unordered" ("nominal" data type). We shall now distinguish between categorical (discrete) data types and real-valued (continuous) data types. Also, for categorical data types we could have unordered or ordered ("ordinal" data type) variables. For an ordered discrete variable $\tilde{x}^d$, we could use Wang and van Ryzin (1981) kernel given by

$$\tilde{l}\left(\tilde{X}_i^d, \tilde{x}^d, \lambda\right) = \begin{cases} 1-\lambda, & \text{if } \tilde{X}_i^d = \tilde{x}^d, \\ \dfrac{(1-\lambda)}{2}\lambda^{|\tilde{X}_i^d - \tilde{x}^d|}, & \text{if } \tilde{X}_i^d \neq \tilde{x}^d. \end{cases}$$

We shall now refer to the unordered kernel defined in Equation 7.2 as $\bar{l}(\cdot)$ so as to keep each kernel type separate notationally speaking. We shall denote the traditional kernels for continuous data types such as the Epanechnikov of Gaussian kernels by $W(\cdot)$.

A generalized product kernel for one continuous, one unordered, and one ordered variable would be defined as follows,

$$K(\cdot) = W(\cdot) \times \bar{l}(\cdot) \times \tilde{l}(\cdot). \tag{7.16}$$

Using such product kernels, we can modify any existing kernel-based method to handle the presence of categorical variables, thereby extending the reach of kernel methods. We define $K_\gamma(X_i, x)$ to be this product, where $\gamma = (h, \lambda)$ is the vector of bandwidths for the continuous and categorical variables.

7.4.1 Kernel Estimation of a Joint Density Defined over Categorical and Continuous Data

Estimating a joint probability/density function defined over mixed data follows naturally using these generalized product kernels. For example, for one unordered discrete variable $\bar{x}^d$ and one continuous variable x^c, our kernel estimator of the PDF would be

$$\hat{f}(\bar{x}^d, x^c) = \frac{1}{nh_{x^c}} \sum_{i=1}^{n} \bar{l}(\bar{X}_i^d, \bar{x}^d) W\left(\frac{X_i^c - x^c}{h_{x^c}}\right).$$

This extends naturally to handle a mix of ordered, unordered, and continuous data (i.e., both quantitative and qualitative data). This estimator is particularly well suited to "sparse data" settings. Li and Racine (2003) demonstrate that

$$\sqrt{nh^p}\left(\hat{f}(z) - f(z) - \hat{h}^2\mathcal{B}_1(z) - \lambda\mathcal{B}_2(z)\right) \to N(0, V(z)) \text{ in distribution,} \tag{7.17}$$

where $\mathcal{B}_1(z) = (1/2)tr\{\nabla^2 f(z)\}[\int W(v)v^2 dv]$, $\mathcal{B}_2(z) = \sum_{x' \in \mathcal{D}, d_{x,x'}=1}[f(x', y) - f(x, y)]$, and $V(z) = f(z)[\int W^2(v)dv]$.

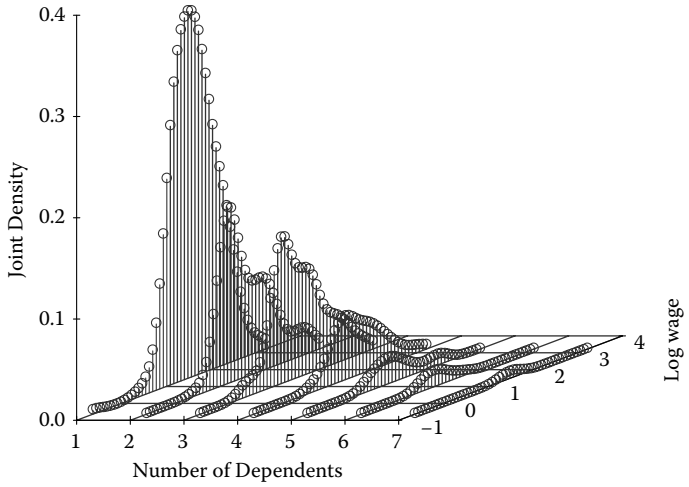

FIGURE 7.1
Nonparametric kernel estimate of a joint density defined over one continuous and one discrete variable.

7.4.1.1 An Application

We consider Wooldridge's (2002) "wage1" dataset having $n = 526$ observations, and model the joint density of two variables, one continuous ("lwage") and one discrete ("numdep"). "lwage" is the logarithm of average hourly earnings for an individual. "numdep" the number of dependents $(0, 1, \ldots)$. We use likelihood cross-validation to obtain the bandwidths, and the resulting estimate is presented in Figure 7.1.

Note that this is indeed a case of "sparse" data for some cells (see Table 7.4), and the traditional approach would require estimation of a nonparametric univariate density function based upon only two observations for the last cell $(c = 6)$.

TABLE 7.4
Summary of the Number of Dependents in the Wooldridge (2002) "wage1" Dataset ("numdep")

	numdep
0	252
1	105
2	99
3	45
4	16
5	7
6	2

7.4.2 Kernel Estimation of a Conditional PDF

Let $f(\cdot)$ and $\mu(\cdot)$ denote the joint and marginal densities of (X, Y) and X, respectively, where we allow Y and X to consist of continuous, unordered, and ordered variables. For what follows we shall refer to Y as a dependent variable (i.e., Y is explained), and to X as covariates (i.e., X is the explanatory variable). We use $\hat{f}$ and $\hat{\mu}$ to denote kernel estimators thereof, and we estimate the conditional density $g(y \mid x) = f(x, y)/\mu(x)$ by

$$\hat{g}(y \mid x) = \frac{\hat{f}(x, y)}{\hat{\mu}(x)}. \tag{7.18}$$

The kernel estimators of the joint and marginal densities $f(x, y)$ and $\mu(x)$ are described in the previous sections; see Hall, Racine, and Li (2004) for details on the theoretical underpinnings of a data-driven method of bandwidth selection for this method.

7.4.2.1 The Presence of Irrelevant Covariates

Hall, Racine, and Li (2004) proposed the estimator defined in Equation 7.18, but choosing appropriate smoothing parameters in this setting can be tricky, not least because plug-in rules take a particularly complex form in the case of mixed data. One difficulty is that there exists no general formula for the optimal smoothing parameters. A much bigger issue is that 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. Hall, Racine, and Li (2004) show that a version of least-squares 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 gives us 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. Cross-validation produces asymptotically optimal smoothing for relevant components, while eliminating irrelevant components by oversmoothing.

Hall, Racine, and Li (2004) demonstrate that, for irrelevant conditioning variables in X, their bandwidths in fact ought to behave exactly the opposite, namely, $h \to \infty$ as $n \to \infty$ for optimal smoothing. The same has been demonstrated for regression as well; see Hall, Li, and Racine (2007) for further details. Note that this result is closely related to the Bayesian results described in detail in Section 7.3.

7.4.3 Kernel Estimation of a Conditional CDF

Li and Racine (2008) propose a nonparametric conditional CDF kernel estimator that admits a mix of discrete and categorical data along with an associated nonparametric conditional quantile estimator. Bandwidth selection for kernel quantile regression remains an open topic of research, and they employ a modification of the conditional PDF-based bandwidth selector proposed by Hall, Racine, and Li (2004).

We use $F(y \mid x)$ to denote the conditional CDF of Y given $X = x$, while $f(x)$ is the marginal density of X. We can estimate $F(y \mid x)$ by

$$\hat{F}(y \mid x) = \frac{n^{-1} \sum_{i=1}^{n} G\left(\frac{y - Y_i}{h_0}\right) K_\gamma(X_i, x)}{\hat{f}(x)}, \tag{7.19}$$

where $G(\cdot)$ is a kernel CDF chosen by the researcher, say, the standard normal CDF, h_0 is the smoothing parameter associated with Y, and $K_\gamma(X_i, x)$ is a product kernel such as that defined in Equation 7.16 where each univariate continuous kernel has been divided by its respective bandwidth for notational simplicity.

Li and Racine (2008) demonstrate that

$$(nh_1 \dots h_q)^{1/2} \left[\tilde{F}(y \mid x) - F(y \mid x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y \mid x) - \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x) \right] \\ \to N(0, V(y \mid x)) \text{ in distribution}, \tag{7.20}$$

where $V(y \mid x) = \kappa^q F(y \mid x)[1 - F(y \mid x)]/\mu(x)$, $B_{1s}(y \mid x) = (1/2)\kappa_2[2F_s(y \mid x) \times \mu_s(x) + \mu(x)F_{ss}(y \mid x)]/\mu(x)$, $B_{2s}(y \mid x) = \mu(x)^{-1} \sum_{z^d \in D} I_s(z^d, x^d)[F(y \mid x^c, z^d) \times \mu(x^c, z^d) - F(y \mid x)\mu(x)]/\mu(x)$, $\kappa = \int W(v)^2 dv$, $\kappa_2 = \int W(v) v^2 dv$, and D is the support of X^d.

7.4.4 Kernel Estimation of a Conditional Quantile

Estimating regression functions is a popular activity for practitioners. Sometimes, however, the regression function is not representative of the impact of the covariates on the dependent variable. For example, when the dependent variable is left (or right) censored, the relationship given by the regression function is distorted. In such cases, conditional quantiles above (or below) the censoring point are robust to the presence of censoring. Furthermore, the conditional quantile function provides a more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function.

Once we can estimate conditional CDFs, estimating conditional quantiles follows naturally. That is, having estimated the conditional CDF we simply invert it at the desired quantile as described below. A conditional αth quantile

of a conditional distribution function $F(\cdot \mid x)$ is defined by ($\alpha \in (0, 1)$)

$$q_\alpha(x) = \inf\{y : F(y \mid x) \geq \alpha\} = F^{-1}(\alpha \mid x).$$

Or equivalently, $F(q_\alpha(x) \mid x) = \alpha$. We can directly estimate the conditional quantile function $q_\alpha(x)$ by inverting the estimated conditional CDF function, i.e.,

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y \mid x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha \mid x).$$

Li and Racine (2008) demonstrate that

$$(nh_1 \ldots h_q)^{1/2}[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)] \to N(0, V_\alpha(x)) \text{ in distribution,} \quad (7.21)$$

where $V_\alpha(x) = \alpha(1-\alpha)\kappa^q/[f^2(q_\alpha(x) \mid x)\mu(x)] \equiv V(q_\alpha(x) \mid x)/f^2(q_\alpha(x) \mid x)$ (since $\alpha = F(q_\alpha(x) \mid x)$).

7.4.5 Binary Choice and Count Data Models

Another application of kernel estimates of PDFs with mixed data involves the estimation of conditional mode models. By way of example, consider some discrete outcome, say $Y \in \mathcal{S} = \{0, 1, \ldots, c-1\}$, which might denote by way of example the number of successful patent applications by firms. We define the conditional mode of $y \mid x$ by

$$m(x) = \max_y g(y \mid x). \quad (7.22)$$

In order to estimate a conditional mode $m(x)$, we need to model the conditional density. Let us call $\hat{m}(x)$ the estimated conditional mode, which is given by

$$\hat{m}(x) = \max_y \hat{g}(y \mid x), \quad (7.23)$$

where $\hat{g}(y \mid x)$ is the kernel estimator of $g(y \mid x)$ defined in Equation 7.18.

7.4.6 Kernel Estimation of Regression Functions

The local constant (Nadaraya 1965; Watson 1964) and local polynomial (Fan 1992) estimators are perhaps the most well-known of all kernel methods. Racine and Li (2004) and Li and Racine (2004) propose local constant and local polynomial estimators of regression functions defined over categorical and continuous data types. To extend these popular estimators so that they can handle both categorical and continuous regressors requires little more than replacing the traditional kernel function with the generalized kernel given in Equation 7.16. That is, the local constant estimator defined in Equation 7.7 would then be

$$\hat{g}(x) = \frac{\sum_{i=1}^n Y_i K_\gamma(X_i, x)}{\sum_{i=1}^n K_\gamma(X_i, x)}. \quad (7.24)$$

Racine and Li (2004) demonstrate that

$$\sqrt{n\hat{h}^p}\left(\hat{g}(x) - g(x) - \hat{B}(\hat{h}, \hat{\lambda})\right) / \sqrt{\hat{\Omega}(x)} \to N(0, 1) \text{ in distribution.} \quad (7.25)$$

See Racine and Li (2004) for further details.

7.5 Summary

We survey recent developments in the kernel estimation of objects defined over categorical and continuous data types. We focus on theoretical underpinnings, and focus first on kernel methods for categorical data only. We pay close attention to recent theoretical work that draws links between kernel methods and Bayesian methods and also highlight the behavior of kernel methods in the presence of irrelevant covariates. Each of these developments leads to kernel estimators that diverge from more traditional kernel methods in a number of ways, and sets the stage for mixed data kernel methods which we briefly discuss. We hope that readers are encouraged to pursue these methods, and draw the readers attention to an R package titled "np" (Hayfield and Racine 2008) that implements a range of the approaches discussed above. A number of relevant examples can also be found in Hayfield and Racine (2008), and we direct the interested reader to the applications contained therein.

References

Aitchison, J., and C. G. G. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* 63(3): 413–420.

Efron, B., and C. Morris. 1973. Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association* 68(341): 117–130.

Fan, J. 1992. Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87: 998–1004.

Hall, P., Q. Li, and J. S. Racine. 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics* 89: 784–789.

Hall, P., J. S. Racine, and Q. Li. 2004. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99(468): 1015–1026.

Hayfield, T., and J. S. Racine. 2008. Nonparametric econometrics: the np package. *Journal of Statistical Software* 27(5). http://www.jstatsoft.org/v27/i05/

Heyde, C. 1997. *Quasi-Likelihood and Its Application*. New York: Springer-Verlag.

Kiefer, N. M., and J. S. Racine. 2009. The smooth colonel meets the reverend. *Journal of Nonparametric Statistics* 21: 521–533.

Li, Q., and J. S. Racine. 2003. Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* 86: 266–292.

Li, Q., and J. S. Racine. 2004. Cross-validated local linear nonparametric regression. *Statistica Sinica* 14(2): 485–512.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Li, Q., and J. S. Racine. 2008. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*. 26(4): 423–434.

Lindley, D. V., and A. F. M. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society* 34: 1–41.

Nadaraya, E. A. 1965. On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* 10: 186–190.

Ouyang, D., Q. Li, and J. S. Racine. 2006. Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics* 18(1): 69–100.

Ouyang, D., Q. Li, and J. S. Racine. 2008. Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory*. 25(1): 1–42.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

Racine, J. S. and Q. Li. 2004. Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119(1): 99–130.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer Series in Statistics.

Wand, M., and B. Ripley, 2008. *KernSmooth: Functions for Kernel Smoothing*. R package version 2.22-22. http://CRAN.R-project.org/package=KernSmooth

Wang, M. C., and J. van Ryzin, 1981. A class of smooth estimators for discrete distributions. *Biometrika* 68: 301–309.

Watson, G. S. 1964. Smooth regression analysis. *Sankhya* 26:(15): 359–372.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.