Smooth Varying-Coefficient Nonparametric Models for Qualitative and Quantitative Data*

Qi Li
Department of Economics, Texas A&M University
College Station, TX 77843-4228, USA,

Jeffrey S. Racine
Department of Economics, McMaster University
Hamilton, ON L8S 4M4, Canada

September 21, 2007
Rough draft not to be circulated or quoted

Abstract

We propose a nonparametric varying coefficient method that admits both qualitative and quantitative regressors. The proposed estimator is exceedingly flexible and has a wide range of potential applications including hierarchical (mixed) settings, small area estimation, and so forth. For example, the method provides for semiparametric models in which practitioners can select a parametric functional form for their model but allow the parameters to change in an unrestricted fashion with respect to, say, a qualitative regressor such as group membership or perhaps with respect to a mix of qualitative and quantitative regressors. However, it also provides a range of semiparametric alternatives that other semiparametric methods cannot match, and can deliver a fully nonparametric estimator should one so desire. A data-driven cross-validatory bandwidth selection method is proposed that can handle both the qualitative and quantitative regressors and can also handle the presence of potentially irrelevant regressors, which can result in finite-sample efficiency gains relative to the conventional frequency estimator that is often found in such settings. Theoretical underpinnings including rates of convergence and asymptotic normality are provided. Monte Carlo simulations are undertaken to assess the method's finite-sample performance relative to the conventional nonparametric frequency estimator, while an empirical application to a seminal dataset is undertaken for illustrative purposes.

Keywords: Categorical regressors, nonparametric smoothing, cross-validation, asymptotic normality.

*Corresponding author: Qi Li, email: qi@econmail.tamu.edu; Tel: 979-845-9954. Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca). Racine would like to thank Tristen Hayfield for his exemplary research assistance.

1 Introduction

The seminal work of Aitchison & Aitken (1976) has spawned a rich literature on the kernel smoothing of discrete (categorical) data, and has made possible numerous advances in the kernel smoothing of datasets comprised of qualitative and quantitative data. One area in which this approach could be particularly helpful is when confronted with data in groups in which potential sparsity can adversely affect one's analysis but which can be alleviated by "borrowing" information from, say, neighboring cells in a prescribed manner. By way of example, one might encounter students grouped in classes grouped in schools or crops grouped in plots of land grouped in farms and so forth. These situations are frequently encountered in mixed model settings such as when conducting small area estimation, estimating multi-level (hierarchical) models, and the like.

The conventional parametric approach to modeling such problems suffers from a number of drawbacks, and this has spawned a rich literature on semiparametric and nonparametric approaches. A variety of less than fully parametric methods have been proposed for estimating these models including semiparametric and nonparametric smoothing spline approaches (Zhang, Lin, Raz & Sowers (1998), Gu & Ma (2005)) and semiparametric kernel-based approaches (Zeger & Diggle (1994)), while a recent *Journal of Multivariate Analysis* (2004) special issue was devoted entirely to such advances. When confronted with a mix of qualitative and qualitative regressors, however, existing semiparametric and nonparametric approaches rely on sample-splitting in order to handle the presence of qualitative regressors, and this can lead to substantial efficiency losses in finite-sample settings. Though existing semiparametric and nonparametric approaches provide the user with much needed flexibility, the method proposed in this paper possesses a number of appealing features not shared by its peers that ought to be quite appealing to practitioners, and does not resort to sample-splitting as it smooths the qualitative regressors in a particular manner as described below.

The estimator we propose is exceedingly flexible. It offers users a semiparametric method wherein the user can select a parametric functional form for their model but allow parameters to change in an unrestricted fashion with respect to group membership. However, it also offers users a fully nonparametric method as well. The rest of the paper proceeds as follows. In sections 2 and 3, theoretical underpinnings including rates of convergence and asymptotic normality are provided along with a data-driven cross-validatory method of bandwidth selection. In Section 4, Monte Carlo simulations are undertaken to assess the finite-sample performance of the proposed method. Section 5 considers an empirical application to a seminal dataset, while Section 6 concludes. Proofs of the main theorems are relegated to the appendices.

2 Kernel Estimation of Varying Coefficient Models

Consider a varying coefficient regression model given by

$$Y_i = X_i'\beta(Z_i) + u_i, \tag{1}$$

where X_i is an p-dimensional vector of regressors; $Z_i = (Z_i^c, Z_i^d)$, where Z_i^c is a continuous regressor of dimension q; and Z_i^d is a discrete regressor of dimension r; the functional form of $\beta(\cdot)$ is not specified, and u_i is an error term satisfying $E(u_i|X_i, Z_i) = 0$.

We use Z_{is}^d to denote the s^{th} component of Z_i^d, and we assume that Z_{is}^d takes c_s ($c_s \geq 2$ is a finite positive integer). We use $\mathcal{S}^d$ to denote the support of Z^d. Also, we allow for one or more of the nonparametric regressors to be irrelevant, where the specific definition of "irrelevant regressors" shall be provided below. Without loss of generality, we assume that the first q_1 ($1 \leq q_1 \leq q$) components of Z^c and the first r_1 ($0 \leq r_1 \leq r$) components of Z^d are "relevant" regressors in the sense defined below. Note that we assume there exists at least one relevant continuous regressor ($q_1 \geq 1$). It can be shown that when all of the continuous regressors are irrelevant, the asymptotic distribution of the cross-validated smoothing parameters will be quite different from the case treated in this paper (i.e., $q_1 \geq 1$). Results for the case in which $q_1 = 0$ cannot be obtained as a special case of those derived below, hence we must treat this case separately and intend do so in future work.

Let $\bar{Z}$ consist of the first q_1 relevant components of Z^c and the first r_1 relevant components of Z^d, and let $\tilde{Z} = Z \setminus \{\bar{Z}\}$ denote the remaining irrelevant components of Z.

Following Hall, Li & Racine (forthcoming) we assume that

$$(Y, X, \bar{Z}) \quad \text{is independent of} \quad \tilde{Z}. \tag{2}$$

Clearly, (2) implies that, for all measurable functions $g(Y, X)$,

$$E(g(Y, X)|Z) = E(g(Y, X)|\bar{Z}). \tag{3}$$

Obviously (2) is a stronger assumption than (3). A weaker condition would be to ask that

$$\text{Conditional on } \bar{Z}, \text{ the variables } \tilde{Z} \text{ and } (Y, X) \text{ are independent.} \tag{4}$$

However, using (4) will cause some formidable technical difficulties in the proofs that follow which we are unable to handle at this time. Therefore, in this paper we will only consider the unconditional independence of (2). Nevertheless, we have also investigated the case of conditional independence as defined by (4) and simulation results reveal that cross-validation can smooth out irrelevant regressors

under either unconditional independence (2) or conditional independence (4).[1]

For an unordered regressor, we suggest using a variant of Aitchison & Aitken's (1976) kernel function defined as

$$l(Z_{is}, z_s, \lambda_s) = \begin{cases} 1, & \text{when } Z_{is} = z_s, \\ \lambda_s, & \text{otherwise.} \end{cases} \tag{5}$$

Note that $\lambda_s = 0$ leads to an indicator function, and $\lambda_s = 1$ gives a uniform weight function. Therefore, the range of λ_s is $[0, 1]$ for all $s = 1, \ldots, r$.

Let $\mathbf{1}(A)$ denote the usual indicator function, which assumes the value 1 if A holds true, and 0 otherwise. Using (5), we can construct a product kernel function given by

$$L(Z_i, x, \lambda) = \prod_{s=1}^{r} l(Z_{is}, z_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{\mathbf{1}(Z_{is} \neq z_s)}. \tag{6}$$

Observe that the kernel weight function we use here does not add up to 1 when $\lambda_s \neq 0$, however, this does not affect the nonparametric estimator $x'\hat{\beta}(z)$ defined in (8) below as the kernel function appears in both the numerator and the denominator of (8), thus the kernel function can be multiplied by any non-zero constant leaving the definition of $\hat{\beta}(z)$ defined below intact.

For the continuous regressors $Z_i^c = (Z_{i1}^c, \ldots, Z_{iq}^c)$ we use the product kernel given by

$$W_h\left(\frac{Z_j^c - Z_i^c}{h}\right) = \prod_{s=1}^{q} \frac{1}{h_s} w\left(\frac{Z_{js}^c - Z_{is}^c}{h_s}\right),$$

where $w(\cdot)$ is a symmetric univariate density function, and where $0 < h_s < \infty$ is the smoothing parameter for z_s^c.

The kernel function for the mixed regressor case $z = (z^c, z^d)$ is simply the product of $W(\cdot)$ and $L(\cdot)$, i.e.,

$$K_{\gamma,ij} = W_h\left(\frac{Z_j^c - Z_i^c}{h}\right) L(Z_j^d, Z_i^d, \lambda), \tag{7}$$

where $\gamma = (h, \lambda) = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$.

We use $\mathcal{S}$ to denote the range assumed by Z_i. For $z \in \mathcal{S}$, from (1) it is fairly straightforward to show that

$$\beta(z) = \left[E(X_i X_i' | Z_i = z)\right]^{-1} E(X_i Y_i | Z_i = z).$$

[1] We do not report these simulations here for space considerations.

Therefore, we estimate $\beta(z)$ by

$$\hat{\beta}(z) = \left[n^{-1} \sum_{i=1}^{n} X_i X_i' K_\gamma(Z_i, z) \right]^{-1} \left[n^{-1} \sum_{i=1}^{n} X_i Y_i K_\gamma(Z_i, z) \right]. \tag{8}$$

When $\lambda_s = 0$ for all $s = 1, \ldots, r$, our estimator reverts back to the conventional approach whereby one uses a frequency estimator to deal with the discrete regressors, while if $\lambda_s = 1$ for some s, then $x'\hat{\beta}(z)$ becomes unrelated to z_s. That is, when $\lambda_s = 1$, z_s is smoothed out from the regression model (it is deemed to be an "irrelevant" qualitative regressor).

We choose $\gamma = (h, \lambda) = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ to minimize[2]

$$CV(\gamma) = \sum_{i=1}^{n} [Y_i - X_i' \hat{\beta}_{-i}(Z_i)]^2 M(Z_i), \tag{9}$$

where $0 \leq M(\cdot) \leq 1$ is a weight function which serves to avoid difficulties caused by dividing by zero, or by the slower convergence rate arising when Z_i lies near the boundary of the support of Z, while

$$\hat{\beta}_{-i}(Z_i) = \left[n^{-1} \sum_{j \neq i}^{n} X_j X_j' K_\gamma(Z_j, Z_i) \right]^{-1} \left[n^{-1} \sum_{j \neq i}^{n} X_j Y_j K_\gamma(Z_j, Z_i) \right] \tag{10}$$

is the leave-one-out kernel estimator of $\beta(Z_i)$.

3 Asymptotic Analysis Allowing for Irrelevant Covariates

We will use $f(x, \bar{z})$ and $\tilde{f}(\tilde{z})$ to denote the joint density of $(X, \bar{Z})$ and $\tilde{Z}$, respectively. Define $\sigma^2(x, \bar{z}) = E(u_i^2 | X_i = x, \bar{Z}_i = \bar{z})$ and let $\mathcal{M}$ denote the support of the weight function $M(\cdot)$. We assume that

> The data are i.i.d. and u_i has finite moments of any order;
>
> $\beta(\bar{z})$, $f(x, \bar{z})$ and $\sigma^2(x, \bar{z})$ have two continuous derivatives (with respect to $\bar{z}^c$);
>
> $M(\cdot)$ is continuous, nonnegative and has compact support; f and $\tilde{f}$ are bounded away from zero for $z = (z^c, z^d) \in \mathcal{S} = \mathcal{S}^c \times \mathcal{S}^d$. (11)

[2]For related work that uses least squares cross-validation for selecting smoothing parameters in a nonparametric regression model with continuous regressors, see Härdle & Marron (1985), and Härdle, Hall & Marron (1988), Härdle, Hall & Marron (1992).

We impose the following conditions on the bandwidth and kernel functions. Define

$$H = (\prod_{s=1}^{q_1} h_s) \prod_{s=q_1+1}^{q} \min(h_s, 1). \tag{12}$$

Letting $0 < \epsilon < 1/(q+4)$ and for some constant $c > 0$, we further assume that

$$\begin{aligned} &n^{\epsilon-1} \leq H \leq n^{-\epsilon}; \quad n^{-c} < h_s < n^c \text{ for all } s = 1, \ldots, q \ ; \\ &\text{the kernel function } w(\cdot) \text{ is a symmetric, compactly supported, Hölder-continuous} \\ &\text{probability density; and } w(0) > w(\delta) \text{ for all } \delta > 0. \end{aligned} \tag{13}$$

The above conditions basically require that each h_s does not converge to zero, or to infinity, too fast, and that $nh_1 \ldots h_{q_1} \to \infty$.

We expect that, as $n \to \infty$, the smoothing parameters associated with the relevant regressors will converge to zero, while those associated with the irrelevant regressors will not. It would be convenient to further assume that $h_s \to 0$ for $s = 1, \ldots, p_1$, and that $\lambda_s \to 0$ for $s = 1, \ldots, q_1$, however, for practical reasons we choose not to assume that the relevant components are known a priori, but rather assume that (15) given below holds. We write $K_{\gamma,ij} = \bar{K}_{\bar{\gamma},ij} \tilde{K}_{\tilde{\gamma},ij}$, where $\bar{\gamma} = (h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{r_1})$, and $\tilde{\gamma} = (h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r)$ so that $\bar{K}$ and $\tilde{K}$ are the product kernel functions associated with the relevant and the irrelevant covariates, respectively.

We define

$$\begin{aligned} \mu_\beta(\bar{z}) &= E[X_j X_j' K_{\gamma,ij} | z_i = z]^{-1} E[X_j X_j' \beta(\bar{z}_j) K_{\gamma,ij} | z_i = z] \\ &= E[X_j X_j' \bar{K}_{\bar{\gamma},ij} | \bar{z}_i = \bar{z}]^{-1} E[X_j X_j' \beta(\bar{z}_j) \bar{K}_{\bar{\gamma},ij} | \bar{z}_i = \bar{z}], \end{aligned} \tag{14}$$

where the second equality comes from the fact that $\tilde{Z}$ is independent of $(X, \bar{Z})$. Therefore, the term related to $\tilde{z}$ cancels out since $E[\tilde{K}_{\tilde{\gamma},ij} | \tilde{z}_i = \tilde{z}]^{-1} E[\tilde{K}_{\tilde{\gamma},ij} | \tilde{z}_i = \tilde{z}] = 1$. Therefore, $\mu_\beta(\bar{z})$ does not depend on $\tilde{z}$, nor does it depend on $(h_{q_1+1}, \ldots, q, \lambda_{r_1+1}, \ldots, \lambda_r)$. We further assume that

$$\begin{aligned} &\textstyle\int_{\text{supp}M} \left[x' \left(\bar{\mu}_\beta(\bar{z}) - \beta(\bar{z})\right)\right]^2 \bar{M}(\bar{z}) f(x, \bar{z}) dx d\bar{z}, \text{ a function of } h_1, \ldots, h_{q_1}, \\ &\text{and } \lambda_1, \ldots, \lambda_{r_1}, \text{ vanishes if and only if all of the smoothing parameters vanish,} \end{aligned} \tag{15}$$

where $\bar{M}$ is a weight function defined in (21) below. In Lemma A.6 in the appendix we show that (13) and (15) imply that as $n \to \infty$, $h_s \to 0$ for $s = 1, \ldots, q_1$ and $\lambda_s \to 0$ for $s = 1, \ldots, r_1$. Therefore, the smoothing parameters associated with the relevant regressors all vanish asymptotically.

Define an indicator function

$$\mathbf{1}_s(v^d, x^d) = \mathbf{1}(v_s^d \neq x_s^d) \prod_{t \neq s, t=1}^{r} \mathbf{1}(v_s^d = x_t^d).$$

Note that $\mathbf{1}_s(v^d, x^d)$ is an indicator function, which equals one if v^d and x^d differ only in their s^{th} component, and zero otherwise.

Let $\int d\bar{z} = \sum_{\bar{z}^d} \int d\bar{z}^c$, let θ_s and θ_{ss} denote the first and second derivatives of θ with respect to z_s^c ($\theta(\bar{z}) = m(\bar{z})$ or $\beta(\bar{z})$), and define

$$
\begin{aligned}
B_{1s}(\bar{z}) &= \left[m_s(z^c, z^d)\beta_s(\bar{z}^c, \bar{z}^d) + (1/2)m(\bar{z}^c, \bar{z}^d)\beta_{ss}(\bar{z}^c, \bar{z}^d) \right] \\
B_{2s}(\bar{z}) &= \sum_{\bar{v}^d \in \bar{S}_d} \mathbf{1}_s(\bar{z}_i^d, \bar{v}^d) m(\bar{z}^c, \bar{v}_d) \left[\beta(\bar{z}^c, \bar{v}^d) - \beta(\bar{z}^c, \bar{z}^d) \right].
\end{aligned} \tag{16}
$$

Then in the appendix we show that the leading term of $CV(\gamma)$ is

$$
\begin{aligned}
&\int \left\{ \left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} \lambda_s B_{2s}(\bar{z}) \right]' m(\bar{z})^{-1} \left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} \lambda_s B_{2s}(\bar{z}) \right] \right\} \bar{M}(\bar{z})\bar{f}(\bar{z})d\bar{z} \\
&\quad + \frac{\kappa^{q_1}}{n h_1 \dots h_{q_1}} \int \delta(\bar{z})\bar{f}(\bar{z})\tilde{f}(\tilde{z})\tilde{R}(\tilde{z})M(z)dz + o_p\left(\zeta_n + (n h_1 \dots h_{q_1})^{-1}\right),
\end{aligned} \tag{17}
$$

where $\zeta_n = \sum_{s=1}^{q_1} h_s^4 + \sum_{s=1}^{r_1} \lambda_s^2$, $\kappa = \int w(v)^2 dv$, $\kappa_2 = \int w(v) v^2 dv$, and where $\tilde{R}(z) = \tilde{R}(z, h_{q_1+1}, \dots, h_q, \lambda_{r_1+1}, \dots, \lambda_r)$ is given by

$$
\tilde{R}(z) = \frac{\nu_2(z)}{[\nu_1(z)]^2}, \tag{18}
$$

where for $l = 1, 2$,

$$
\nu_l(z) = E\left(\left[\prod_{s=q_1+1}^{q} h_s^{-1} w\left(\frac{z_{is}^{\mathrm{c}} - z_s^{\mathrm{c}}}{h_s} \right) \prod_{s=r_1+1}^{r} \lambda_s^{\mathbf{1}(z_{is}^d \neq z_s^d)} \right]^l \right). \tag{19}
$$

In (17) the irrelevant regressor $\tilde{z}$ appears in $\tilde{R}(z)$. By Hölder's inequality, $\tilde{R}(z) \geq 1$ for all choices of z, $h_{p_1+1}, \dots, h_p$, and $\lambda_{q_1+1}, \dots, \lambda_q$. Also, $\tilde{R}(z) \to 1$ as $h_s \to \infty$ ($q_1 + 1 \leq s \leq q$) and $\lambda_s \to 1$ ($r_1 + 1 \leq s \leq r$). Therefore, in order to minimize (17), one needs to select h_s ($s = q_1 + 1, \dots, q$) and λ_s ($s = r_1 + 1, \dots, r$) to minimize $\tilde{R}(z)$. In fact, we show that the only smoothing parameter values for which $\tilde{R}(z, h_{q_1+1}, \dots, h_q, \lambda_{r_1+1}, \dots, \lambda_r) = 1$ are $h_s = \infty$ for $q_1 + 1 \leq s \leq q$, and $\lambda_s = 1$ for $r_1 + 1 \leq s \leq r$. To see this, let us define $Z_n = \prod_{s=q_1+1}^{q} w\left(\frac{z_s^{\mathrm{c}} - Z_{is}^{\mathrm{c}}}{h_s} \right) \prod_{s=r_1+1}^{r} \lambda_s^{\mathbf{1}(z_s^{\mathrm{d}} \neq Z_{is}^{\mathrm{d}})}$. If at least one h_s is finite (for $q_1 + 1 \leq s \leq q$), or one $\lambda_s < 1$ (for $r_1 + 1 \leq s \leq r$), then by (13) ($w(0) > w(\delta)$ for all $\delta > 0$) we know that $\mathrm{Var}(Z_n) = E[Z_n^2] - [E(Z_n)]^2 > 0$ so that $R = E[Z_n^2]/[E(Z_n)]^2 > 1$. Only when, in the definition of Z_n, all $h_s = \infty$ and all $\lambda_s = 1$, do we have $Z_n \equiv w(0)^{q-q_1}$ (a constant) and $\mathrm{Var}(Z_n) = 0$ so that $\tilde{R}(z) = 1$ only in this case.

Therefore, in order to minimize (17), the smoothing parameters corresponding to the irrelevant regressors must all converge to their upper extremities, so that $\tilde{R}(z) \to 1$ as $n \to \infty$ for all $z \in \mathcal{S}$. Thus, the irrelevant components are asymptotically smoothed out.

To analyze the behavior of smoothing parameters associated with the relevant regressors, we replace $\tilde{R}(z)$ by 1 in (17), thus the first term on the right-hand-side of (17) becomes

$$\frac{\kappa^{q_1}}{nh_1 \dots h_{q_1}} \int \sigma^2(x,\bar{z}) m(\bar{z}) \bar{M}(\bar{z}) f(x,\bar{z}) d\bar{z} dx \tag{20}$$

where

$$\bar{M}(\bar{z}) = \int \tilde{f}(\tilde{z}) M(z) d\tilde{z}, \tag{21}$$

where, again, we define $\int d\tilde{z} = \sum_{\tilde{z}^d} \int d\tilde{z}^c$.

Next, defining $a_s = h_s n^{1/(q_1+4)}$ and $b_s = \lambda_s n^{2/(q_1+4)}$, then (17) (with (20) as its first term since $\tilde{R}(z) \to 1$) becomes $n^{-4/(q_1+4)} \bar{\chi}(a_1, \dots, a_{q_1}, b_1, \dots, b_{r_1})$, where

$$\begin{aligned} \bar{\chi}(a_1, \dots, b_{r_1}) = {} & \frac{\kappa^{q_1}}{a_1 \dots a_{q_1}} \int \delta(\bar{z}) m(\bar{z}) \bar{f}(\bar{z}) \bar{M}(\bar{z}) f(x,\bar{z}) d\bar{z} dx \\ & + \int \left\{ \left[\sum_{s=1}^{q_1} a_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} b_s B_{2s}(\bar{z}) \right]' m(\bar{z}_i)^{-1} \left[\sum_{s=1}^{q_1} a_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} b_s B_{2s}(\bar{z}) \right] \right\} \bar{M}(\bar{z}) \bar{f}(\bar{z}) d\bar{z}. \end{aligned} \tag{22}$$

Let $a_1^0, \dots, a_{p_1}^0, b_1^0, \dots, b_{q_1}^0$ denote values of $a_1, \dots, a_{p_1}, b_1, \dots, b_{q_1}$ that minimize $\bar{\chi}$ subject to each of them being nonnegative. We require that

$$\text{Each } a_s^0 \text{ is positive and each } b_s^0 \text{ nonnegative, all are finite and uniquely defined.} \tag{23}$$

Li & Zhou (2005) give simple necessary and sufficient conditions that ensure that (23) holds true.

Theorem 3.1. *Assume conditions* (11), (13), (15), *and* (23) *hold, and let* $\hat{h}_1, \dots, \hat{h}_q$, $\hat{\lambda}_1, \dots, \hat{\lambda}_r$ *denote the smoothing parameters that minimize* $CV(\gamma)$. *Then*

$$\begin{aligned} & n^{1/(p_1+4)} \hat{h}_s \to a_s^0 \textit{ in probability for } 1 \le s \le q_1, \\ & P(\hat{h}_s > C) \to 1 \textit{ for } q_1 + 1 \le s \le q \textit{ and for all } C > 0, \\ & n^{2/(q_1+4)} \hat{\lambda}_s \to b_s^0 \textit{ in probability for } 1 \le s \le r_1, \\ & \hat{\lambda}_s \to 1 \textit{ in probability for } r_1 + 1 \le s \le r, \end{aligned} \tag{24}$$

and $n^{4/(q_1+4)} CV(\hat{h}_1, \dots, \hat{h}_q, \hat{\lambda}_1, \dots, \hat{\lambda}_r) \to \inf \bar{\chi}$ *in probability.*

The proof of Theorem 3.1 is given in the appendix. From Theorem 3.1 one can easily obtain the following result.

Theorem 3.2. *Under the same conditions given in Theorem 3.1, then*

$$\sqrt{nh_1 \dots h_{q_1}} \left(\hat{\beta}(z) - \beta(\bar{z}) - \sum_{s=1}^{q_1} \hat{h}_s B_{1s}(\bar{z}) - \sum_{s=1}^{r_1} \hat{\lambda}_s B_{2s}(\bar{z}) \right) / \sqrt{\hat{\Omega}(\bar{z})} \to N(0,1) \text{ in distribution,}$$

where $\hat{\Omega}(z) = []$

4 Finite-Sample Performance

In this section we outline a modest Monte Carlo simulation designed to highlight the finite-sample behavior of the proposed method. For what follows, we shall consider the behavior of two estimators, namely the proposed kernel method and the conventional frequency-based kernel method that breaks the data into subsets.

We simulate data from

$$Y_i = \beta_{0j} + \beta_{1ji} \sin(X_{i1}) + \beta_2 X_{i2} + \epsilon_i, \qquad i = 1, \dots, n, \qquad j = 0, \dots, c-1,$$

where X_{i1} and X_{i2} are $U[-2\pi, 2\pi]$, $\epsilon_i \sim N(0,1)$,

$$\begin{aligned} \beta_{0j} =& \beta_0 + \eta_{0j}, \qquad \eta_{0j} \sim N(0,1), \qquad \beta_0 = 1, \qquad j = 0, \dots, c-1, \\ \beta_{1ji} =& \beta_1 + (Z_i^c)^2 + \eta_{1j}, \qquad \eta_{1ji} \sim N(0,1), \qquad \beta_1 = 1, \qquad j = 0, \dots, c-1, \end{aligned}$$

with $Z_{i1}^c \sim N(0,1)$.

We draw c subsets of size n/c ($n > c$ and divides evenly) so that if, say, $n = 100$ and, say, $c = 2$, then we would have two subsets consisting of 50 observations for which $Y_i = \beta_{00} + \beta_{10i} \sin(X_{i1}) + \beta_2 X_{i2} + \epsilon_i$ and 50 observations for which $Y_i = \beta_{01} + \beta_{11i} \sin(X_{i1}) + \beta_2 X_{i2} + \epsilon_i$. Next we let Z_{i1}^d denote 'group membership', i.e., if, say, $n = 100$ and, say, $c = 2$, then for the 50 observations for which $Y_i = \beta_{00} + \beta_{10} \sin(X_{i1}) + \beta_2 X_{i2} + \epsilon_i$ we set $Z_{i1}^d = 0$, while for the 50 observations for which $Y_i = \beta_{01} + \beta_{11} \sin(X_{i1}) + \beta_2 X_{i2} + \epsilon_i$ we set $Z_{i1}^d = 1$. Finally, we generate another discrete regressor $Z_{i2}^d \in \{0, \dots, c-1\}$ that is uncorrelated with Y_i, i.e., is "irrelevant" though we do not presume this is known a priori. One can interpret Z_1^d as a discrete covariate (say, 'group 1 membership') where the model changes with respect to Z_1^d, while Z_2^d can be interpreted as a discrete covariate (say, 'group 2 membership') where there is no variation in the model with respect to this group.

In other words, the true data generating process is a function of X_1, X_2, Z_1^c and Z_1^d only, namely

$$Y_i = \gamma_0(Z_{i1}^d) + \gamma_1(Z_{i1}^d, Z_{1i}^c) \sin(X_{i1}) + \gamma_2 X_{i2} + \epsilon_i.$$

However, Z_2^d is an irrelevant covariate, but this is not known a priori hence the user includes all regressors in the specification. The nonparametric model is of the form

$$Y_i = \gamma_0(Z_{i1}^d, Z_{i2}^d, Z_{i1}^c) + \gamma_1(Z_{i1}^c, Z_{i1}^d, Z_{i2}^d)X_{i1} + \gamma_2(Z_{i1}^c, Z_{i1}^d, Z_{i2}^d)X_{i2} + u_i.$$

We conduct $M = 1,000$ Monte Carlo replications from this DGP, estimate each model, consider settings having 4, 16, and 25 cells of data and consider samples of size $n = 100, 200, 300, 400, 500$.

In Table 1 we report the relative median mean square error (MSE) for the smooth and frequency-based kernel approaches (i.e., the ratio of the median MSE of the conventional sample-splitting approach to the median MSE of the proposed approach). Numbers greater than one indicate a loss of efficiency relative to the proposed method..

Table 1: Relative median MSE

	$c = 2$	$c = 4$	$c = 5$
100	1.37	2.74	3.54
200	1.21	2.06	2.61
300	1.12	1.67	2.15
400	1.09	1.51	1.86
500	1.06	1.39	1.70

A quick scan of Table 1 reveals that the conventional nonparametric approach whereby one breaks data into subsets results in substantial efficiency losses that worsen as the number of subsets increases and/or the sample size falls.

Next we consider the performance of the cross-validated bandwidths for Z_1^d ($\hat{\lambda}_1$), Z_2^d ($\hat{\lambda}_2$), and Z_1^c ($\hat{h}$). Tables 2 through 4 summarize the median bandwidths over the $M = 1,000$ Monte Carlo replications.

Tables 2 through 4 reveal that the bandwidths for the relevant regressors ($\hat{\lambda}_1$ and $\hat{h}$) behave as expected converging to zero as n increases, while that for the irrelevant regressor ($\hat{\lambda}_2$) converges instead to its upper bound value of one in probability as n increases.

This modest Monte Carlo highlights the fact that the proposed method that smooths both the discrete and continuous covariates is more efficient that the conventional frequency-based approach. It also illustrates how cross-validated bandwidth selection delivers appropriate bandwidths for both relevant and irrelevant covariates. We now turn to a modest application that underscores the flexibility of the proposed method.

Table 2: Median bandwidths ($c = 2$)

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{h}$
100	0.47	0.90	0.77
200	0.43	1.00	0.77
300	0.36	1.00	0.78
400	0.32	1.00	0.80
500	0.30	1.00	0.78

Table 3: Median bandwidths ($c = 4$)

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{h}$
100	0.39	0.83	0.83
200	0.30	1.00	0.81
300	0.28	1.00	0.85
400	0.24	1.00	0.90
500	0.21	1.00	0.91

Table 4: Median bandwidths ($c = 5$)

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{h}$
100	0.34	0.83	0.87
200	0.29	1.00	0.86
300	0.25	1.00	0.84
400	0.21	1.00	0.84
500	0.19	1.00	0.86

5 Application

As noted, the proposed mixed data estimator is exceedingly flexible. For what follows, we consider using the method in a multi-level model setting. We consider a well-known dataset which was analyzed by Kreft & de Leeuw (1998, ch. 4). The data consists of a sample from the NELS-88 data in which student achievement at 23 randomly selected schools is analyzed, and interest lies mainly on within school dispersions and correlations. The response variable is a student's math achievement ('MathAchievement'). In addition to the response variable, we make use of the following regressors, namely the number of hours of homework done per week ('HomeWork') and a school identifier ('Schid'). The sample contains $n = 519$ observations.

We first replicate the example found in Kreft & de Leeuw (1998, Section 4.2.4, Table 4.4) and allow for a random slope for "HomeWork" in the parametric model and obtain the "fixed effects"

parameters (average values) given by

$$\text{MathAchievement} = 46.3 + 2.0 \times \text{HomeWork}. \tag{25}$$

Next we consider a semiparametric model letting $X = (\text{HomeWork})$ and $Z = (\text{schid})$ treating the school identifier as a qualitative regressor.[3] The semiparametric fixed effects parameters are given by

$$\text{MathAchievement} = 47.2 + 2.1 \times \text{HomeWork}. \tag{26}$$

The bandwidth for schid is $\hat{\lambda} = 0.001$. Comparing (25) and (26), it can be seen that the parametric and semiparametric fixed effects estimates are indistinguishable from one another. However, this might be driven by the presumed linearity of the underlying relationship, which we now investigate.

One of the strengths of the proposed method is that if we incorporate X in the conditioning set Z, then the method is capable of modeling nonlinearities in X. However, if the underlying relationship is linear in X then cross-validation ought to oversmooth X delivering a model that is approximately linear in X. That is, the proposed method can be fully nonparametric but will revert to a semiparametric model such as (26) if the semiparametric model is appropriate.

To investigate this issue further we next consider a nonparametric model, and let $X = (\text{HomeWork})$ and $Z = (\text{HomeWork}, \text{schid})$ treating hours of homework as a continuous regressor and the school identifier as a qualitative one (so now $q_1 = 1$). The nonparametric fixed effects parameters are given by

$$\text{MathAchievement} = 47.3 + 2.0 \times \text{HomeWork}, \tag{27}$$

and it is evident that (25), (26), and (27) are approximately equivalent.

The relationship between MathAchievement and HomeWork turns out to be approximately linear as can be seen by examining Table 5 which summarizes the bandwidths for the nonparametric model.

Table 5: Bandwidth summary underlying (27).

Name	HomeWork	factor(Schid)
Bandwidth	$\hat{h} = 3.255$	$\hat{\lambda} = 0.001$

Table 5 reveals that the bandwidth for the regressor HomeWork is $\hat{h} = 3.255 = 7.674\hat{\sigma}n^{-1/5}$ where

[3]The astute reader will note that, in this specification, $q_1 = 0$ and that we have not in fact covered this case. We beg the reader's indulgence as we consider the case for which $q_1 = 1$ in the case that immediately follows.

$\hat{\sigma}$ is the standard deviation of HomeWork, which represents a very large degree of oversmoothing and indicates that the model is approximately linear in HomeWork. We get virtually identical results for the semiparametric model that does not incorporate HomeWork in Z, as can be seen by comparing (26) and (27). For this reason, we get almost identical results for the fixed effects whether using the semiparametric, nonparametric, or the parametric approach. Next we examine the random effects for the two models, which are summarized in Table 6.

Table 6: Summary of random effects for the parametric and nonparametric models.

Parametric Random Effects

Groups	Name	Variance	Corr
Schid	(Intercept)	62.393	
	HomeWork	17.703	-0.829
Residual		53.297	
number of obs: 519, groups: Schid, 23			

Nonparametric Random Effects

Groups	Name	Variance	Corr
Schid	(Intercept)	65.775	
	HomeWork	14.879	-0.746
Residual		48.132	
number of obs: 519, groups: Schid, 23			

Table 6 reveals that the the random effects variance appears to be comparable for the intercept (within 5% of one another) but the random effects variance for HomeWork differs by four times as much (19%). The differences among the random effects may arise from the fact that the parametric model presumes normality for the distribution of the random effects which may not hold in this setting. To further investigate this issue, we conducted a Shapiro-Wilks test for normality for the random coefficients for the parametric model, which rejects normality for the random coefficient on HomeWork ($P = 0.015$) but not for the random intercept ($P = 0.352$). This suggests that the parametric model is misspecified as it presumes that the random effects coefficients are drawn from normal distributions, which is rejected by the data for the variable of interest (HomeWork).

We hope that this modest application to a well-known dataset will encourage practioners to investigate the proposed method in settings such as that described above.

6 Summary

In this paper we propose a novel varying coefficient model that is capable of handling the mix of qualitative and quantitative regressors commonly encountered in applied work. A data-driven bandwidth selection method is proposed, theoretical underpinnings are provided, a Monte Carlo experiment is undertaken to examine the finite-sample behavior of the proposed method, while an application to a popular dataset demonstrates the utility of the proposed method in applied settings. Results for the case in which all elements of Z are qualtitative cannot be obtained as a special case of those derived herein, hence we must treat this case separately and intend do so in future work.

References

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

Gu, C. & Ma, P. (2005), 'Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection', *Journal of Computational and Graphical Statistics* **14**, 485–504.

Hall, P., Li, Q. & Racine, J. S. (forthcoming), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors', *The Review of Economics and Statistics* .

Härdle, W., Hall, P. & Marron, J. S. (1988), 'How far are automatically chosen regression smoothing parameters from their optimum?', *Journal of The American Statistical Association* **83**, 86–101.

Härdle, W., Hall, P. & Marron, J. S. (1992), 'Regression smoothing parameters that are not far from their optimum', *Journal of the American Statistical Association* **87**, 227–233.

Härdle, W. & Marron, J. (1985), 'Optimal bandwidth selection in nonparametric regression function estimation', *The Annals of Statistics* **13**, 1465–1481.

Journal of Multivariate Analysis (2004), Vol. 91, Academic Press, Inc., Orlando, FL, USA. Special issue on semiparametric and nonparametric mixed models.

Kreft, I. & de Leeuw, J. (1998), *Introducing Multilevel Modeling*, Sage.

Li, Q. & Zhou, J. (2005), 'The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models', *Econometric Theory* **21**(5), 1017–1025.

Masry, E. (1996), 'Multivariate regression estimation: local polynomial fitting for time series', *Stochastic Processes and Their Applications* **65**, 81–101.

Zeger, S. L. & Diggle, P. J. (1994), 'Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters', *Biometrics* **50**, 689–699.

Zhang, D., Lin, X., Raz, J. & Sowers, M. (1998), 'Semiparametric stochastic mixed models for longitudinal data', *Journal of the American Statistical Association* **93**(442), 710–719.

A Proofs of Theorem 3.1 and Theorem 3.2

A.1 Preliminaries

The proof of Theorem 3.1 is quite tedious. Therefore, it is necessary to introduce some short-hand notation and preliminary manipulations in order to simplify the derivations that follow. For the reader's convenience we list most of the notation used in this appendix directly below.

1. We will use β_i to denote $\beta(Z_i)$ and $\hat{\beta}_{-i}$ to denote $\hat{\beta}_{-i}(Z_i)$.

2. We define $\sum_i = \sum_{i=1}^n$, $\sum\sum_{j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n$, $\sum\sum\sum_{l\neq j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n \sum_{l=1,l\neq i,l\neq j}^n$.

3. We write $A_n = B_n + (s.o.)$ to denote the fact that B_n is the leading term of A_n, where $(s.o.)$ denotes terms that have orders smaller than B_n. $A_i = B_i + (s.o.)$ always means that $n^{-1}\sum_i A_i = n^{-1}\sum_i B_i + (s.o.)$, and $A_{ij} = B_{ij} + (s.o.)$ means that $n^{-2}\sum_i\sum_j A_{ij} = n^{-2}\sum_i\sum_j B_{ij} + (s.o.)$. Also, we write $A_n \sim B_n$ to mean that A_n and B_n have the same order of magnitude in probability.

4. For notational simplicity we often ignore the difference between n^{-1} and $(n-1)^{-1}$ simply because this will have no effect on the asymptotic analysis.

Proof of Theorem 3.1. Using (9) and $Y_i = X_i'\beta_i + u_i$, we have (where $\beta_i = \beta(z_i)$, $\hat{\beta}_{-i} = \hat{\beta}_{-i}(z_i)$ and $M_i = M(z_i)$)

$$\begin{aligned} CV(\gamma) &= \frac{1}{n}\sum_i [Y_i - X_i'\hat{\beta}_{-i}]^2 M_i \\ &= \frac{1}{n}\sum_i [X_i'(\beta_i - \hat{\beta}_{-i})]^2 M_i + \frac{2}{n}\sum_i u_i X_i'(\beta_i - \hat{\beta}_{-i})M_i + n^{-1}\sum_i u_i^2 M_i. \end{aligned} \tag{A.1}$$

Below we obtain the leading terms of $CV(\lambda)$. We use $CV_0(\lambda)$ to denote the first two terms on the right-hand-side of (A.1). Minimizing $CV(\gamma)$ over $\gamma = (h, \lambda)$ is equivalent to minimizing $CV_0(\gamma)$ (the first two terms of $CV(\gamma)$) as $n^{-1}\sum_i u_i^2$ does not depend on λ, where

$$CV_0(\gamma) \stackrel{def}{=} n^{-1}\sum_i [X_i'(\beta_i - \hat{\beta}_{-i})]^2 M_i + 2n^{-1}\sum_i u_i X_i'(\beta_i - \hat{\beta}_{-i})M_i = CV_{0,1} + CV_{0,2}, \tag{A.2}$$

where the definitions of $CV_{0,1}$ and $CV_{0,2}$ should be apparent.

Replacing $Y_j = X_j\beta_j + u_j = X_j'\beta_i + X_j'(\beta_j - \beta_i) + u_j$ in (10), we get

$$\begin{aligned}\hat{\beta}_{-i} &= \beta_i + \left[\hat{A}(z_i)\right]^{-1}\left[n^{-1}\sum_{j\neq i} X_jX_j'(\beta_j-\beta_i)K_{\gamma,ij} + n^{-1}\sum_{j\neq i} X_ju_jK_{\gamma,ij}\right] \\ &= \beta_i + \left[\hat{A}(z_i)\right]^{-1}\left[\hat{B}_i + \hat{C}_i\right], \end{aligned} \tag{A.3}$$

where $\hat{A}(z_i) = n^{-1}\sum_{j\neq i}^n X_jX_j'K_{\gamma,ij}$, $\hat{B}_i = n^{-1}\sum_{j\neq i}^n X_jX_j'(\beta_j-\beta_i)K_{\gamma,ij}$ and $\hat{C}_i = n^{-1}\sum_{j\neq i}^n X_ju_jK_{\gamma,ij}$.

Substituting (A.3) into $CV_{0,1}$ we get

$$\begin{aligned} CV_{0,1} &= n^{-1}\sum_i \left\{X_i'\hat{A}(z_i)^{-1}\left[\hat{B}_i+\hat{C}_i\right]\right\}^2 M_i \\ &= n^{-1}\sum_i \left[X_i'\hat{A}(z_i)^{-1}\hat{B}_i\right]^2 + n^{-1}\sum_i \left[X_i'\hat{A}(z_i)^{-1}\hat{C}_i\right]^2 M_i \\ &\quad + 2n^{-1}\sum_i \left[X_i'\hat{A}(z_i)^{-1}\hat{B}_i\right]\left[X_i'\hat{A}(z_i)^{-1}\hat{C}_i\right] M_i \\ &= CV_1 + CV_2 + 2CV_3. \end{aligned}$$

In lemmas A.1 through A.3 below we show, uniformly in $(x,z)\in S$, and in $\gamma\in\Gamma$, that ($B_{ls}(\cdot)$ is defined in (16))

$$\begin{aligned} CV_1 &= \int\left\{\left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1}\lambda_s B_{2s}(\bar{z})\right]' m(\bar{z})^{-1}\right. \\ &\quad \left.\times\left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1}\lambda_s B_{2s}(\bar{z})\right]\right\}\bar{f}(\bar{z})\bar{M}(\bar{z})d\bar{z} + o_p\left(\zeta_n^2 + (nh_1\ldots h_{q_1})^{-1}\right), \end{aligned} \tag{A.4}$$

$$CV_2 = \frac{\kappa^{q_1}}{nh_1\ldots h_{q_1}}\int \bar{f}(\bar{z})\tilde{f}(\tilde{z})\delta(\bar{z})\nu_2(\tilde{z})\nu_1(\tilde{z})^{-2}M(z)dz, \tag{A.5}$$

$$CV_3 = o_p\left(\zeta_n^2 + (nh_1\ldots h_{q_1})^{-1}\right), \tag{A.6}$$

where $\delta(\bar{z}_i,\bar{z}_j) = E[(x_i'\mu(\bar{z}_i)^{-1}x_j)^2\sigma^2(x_j,\bar{z}_j)|\bar{z}_i,\bar{z}_j]$ and $\zeta_n = \sum_{s=1}^{q_1} h_s^4 + \sum_{s=1}^{r_1}\lambda_s^2$. Also in Lemma A.4 we show that

$$CV_{0,2} = o_p(\zeta_n + (nh_1\ldots h_{q_1})^{-1}). \tag{A.7}$$

Hence, the leading terms of $CV(\gamma)$ are given by $CV_1 + CV_2$.

In (A.5) the irrelevant regressor $\tilde{z}$ appears in $\tilde{R} \overset{def}{=} \nu_2(\tilde{z})/\nu_1(\tilde{z})^2$. By Hölder's inequality, $\tilde{R}\geq 1$ for all choices of $\tilde{z}$, $h_{q_1+1},\ldots,h_q$, and $\lambda_{r_1+1},\ldots,\lambda_r$. Also, $\tilde{R}\to 1$ as $h_s\to\infty$ ($q_1+1\leq s\leq q$) and $\lambda_s\to 1$ ($r_1+1\leq s\leq r$). Therefore, in order to minimize (A.5), one needs to select h_s ($s=q_1+1,\ldots,q$) and

λ_s $(s = r_1+1, \ldots, r)$ to minimize $\tilde{R}$. In fact, we show below that the only smoothing parameter values for which $\tilde{R} = \tilde{R}(\tilde{z}, h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r) = 1$ are $h_s = \infty$ for $q_1 + 1 \leq s \leq q$, and $\lambda_s = 1$ for $r_1 + 1 \leq s \leq r$. To see this, let us define $A_n(\tilde{z}) = \prod_{s=q_1+1}^{q} w\left(\frac{\tilde{z}_s^c - \tilde{z}_{is}^c}{h_s}\right) \prod_{s=r_1+1}^{r} \lambda_s^{\mathbf{1}(\tilde{z}_s^d \neq \tilde{z}_{is}^d)}$. If at least one h_s is finite (for $q_1+1 \leq s \leq q$), or one $\lambda_s < 1$ (for $r_1+1 \leq s \leq r$), then by (13) that $w(0) > w(\delta)$ for all $\delta > 0$, we know that $\mathrm{Var}(A_n) = E[A_n^2] - [E(A_n)]^2 > 0$ so that $\tilde{R} = E[A_n^2]/[E(A_n)]^2 > 1$. Only when, in the definition of A_n, all $h_s = \infty$ and all $\lambda_s = 1$, do we have $A_n \equiv w(0)^{q-q_1}$ (a constant) and $\mathrm{Var}(A_n) = 0$ so that $\tilde{R} = 1$ only in this case. □

Lemma A.1. *Equation* (A.4) *holds true.*

Proof. By Lemma A.5 we know that the leading term of $\hat{A}(z_i)^{-1}$ is $\mu(z_i)^{-1}$, where $\mu(z_i)^{-1} = m(\bar{z}_i)E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i = \tilde{z}]$ and $m(\bar{z}) = E[X_jX_j'|\bar{z}_j = \bar{z}]\bar{f}(\bar{z})$. Define CV_1^0 by replacing $\hat{A}(z_i)^{-1}$ in CV_1 by $\mu(z_i)^{-1}$. Then we have $CV_1 = CV_1^0 + (s.o.)$. Hence, we only need to consider CV_1^0 which we write as

$$\begin{aligned}
CV_1^0 &= n^{-1}\sum_i \left[X_i'\mu(z_i)^{-1}\hat{B}_i\right]^2 \\
&= n^{-3}\sum_i \sum_{j\neq i} \sum_{l\neq i} X_i'\mu(z_i)^{-1}X_jX_j'(\beta_j - \beta_i)K_{\gamma,ij}X_i'\mu(z_i)^{-1}X_lX_l'(\beta_l - \beta_i)K_{\gamma,il} \\
&= n^{-3}\sum_i \sum_{j\neq i} \left[X_i'\mu(z_i)^{-1}X_jX_j'(\beta_j - \beta_i)K_{\gamma,ij}\right]^2 \\
&\quad + n^{-3}\sum_i \sum_{j\neq i} \sum_{l\neq i, l\neq j} X_i'\mu(z_i)^{-1}X_jX_j'(\beta_j - \beta_i)K_{\gamma,ij}X_i'\mu(z_i)^{-1}X_lX_l'(\beta_l - \beta_i)K_{\gamma,il} \\
&= CV_{1,1} + CV_{1,2}
\end{aligned}$$

Noting that $\mu(z_i) = m(\bar{z}_i)\nu_1(\tilde{z})$ and $\nu_2(\tilde{z}) = E[\tilde{K}_{\tilde{\gamma},ij}^2|\tilde{z}_i]$, it is fairly straightforward to show that

$$\begin{aligned}
E\left[|CV_{1,1}|\right] &= n^{-3}n(n-1)E\left[X_i'm(\bar{z}_i)^{-1}X_jX_j'(\beta_j - \beta_i)^2\bar{K}_{\bar{\gamma},ij}\right]^2 E\left\{\tilde{K}_{\tilde{\gamma},ij}^2/(E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i])^2\right\} \\
&\leq C(nh_1\ldots h_{q_1})^{-1}\left(\sum_{i=1}^{q_1} h_s^2 + \sum_{s=1}^{r_1}\lambda_s\right) = o\left((nh_1\ldots h_{q_1})^{-1}\right).
\end{aligned}$$

since $E\{[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i]^2/(E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i])^2\} \leq 1$. Hence, $CV_{1,1} = O_p((nh_1\ldots h_{q_1})^{-1})$.

Next, by the U-statistic H-decomposition, one can show that $CV_{1,2} = E[CV_{1,2}] + (s.o.)$. Before evaluating $E[CV_{1,2}]$, we first compute an intermediate quantity. Noting that $\mu(z) = m(\bar{z})\nu_1(\tilde{z})$ and

$m(\bar{z}) = E[X_j X_j' | \bar{z}_j = \bar{z}]\bar{f}(\bar{z})$, we have

$$
\begin{aligned}
&E\left[x_i'\mu(z_i)^{-1}x_j x_j'(\beta_j - \beta_i)K_{\gamma,ij}|x_i, z_i\right] = X_i' m(z_i)^{-1} E\left[E(x_j x_j'|\bar{z}_j)(\beta_j - \beta_i)\bar{K}_{\gamma,ij}|x_i, z_i\right] \\
&= x_i' m(z_i)^{-1} \int \bar{f}(\bar{z}_j) E(x_j x_j'|\bar{z}_j)(\beta_j - \beta_i)\bar{K}_{\gamma,ij} dz_j \\
&= x_i' m(\bar{z}_i)^{-1} \sum_{\bar{z}^d \in \bar{S}_d} \int m(\bar{z}_i^c + hv, \bar{z}^d)(\beta(\bar{z}_i^c + hv, \bar{z}^d) - \beta(\bar{z}_i^c, \bar{z}_i^d))W(v)L(\bar{z}_i^d, \bar{z}^d, \lambda)dv \\
&= x_i' m(\bar{z}_i)^{-1} \left\{ \sum_{s=1}^{q_1} h_s^2 \left[m_s(z_i^c, z_i^d)\beta_s(\bar{z}_i^c, \bar{z}_i^d) + (1/2)m(\bar{z}_i^c, \bar{z}_i^d)\beta_{ss}(\bar{z}_i^c, \bar{z}_i^d) \right] \right. \\
&\qquad \left. + \sum_{s=1}^{r_1} \lambda_s \sum_{\bar{v}^d \in \bar{S}_d} I_s(\bar{z}_i^d, \bar{v}^d) m(\bar{z}_i^c, \bar{v}_d) \left[\beta(\bar{z}_i^c, \bar{v}^d) - \beta(\bar{z}_i^c, \bar{z}_i^d) \right] \right\} + o_p(\zeta_n) \\
&\equiv x_i' m(\bar{z}_i)^{-1} \left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}_i) + \sum_{s=1}^{r_1} \lambda_s B_{2s}(\bar{z}_i) \right] + o_p(\zeta_n) \qquad \text{(A.8)}
\end{aligned}
$$

uniformly in $\gamma \in \Gamma$.

Using (A.8) and letting $\xi(\bar{z}) = E[X_i X_i' | \bar{z}_j = \bar{z}]$ (then $m(\bar{z}) = \xi(\bar{z})\bar{f}(\bar{z})$), we immediately obtain

$$
\begin{aligned}
E[CV_{1,2}] &= E\left\{ \left[\sum_{s=1}^{q_1} h_s^2 B_{1s,i} + \sum_{s=1}^{r_1} \lambda_s B_{2s,i} \right]' m(\bar{z}_i)^{-1} \xi(\bar{z}_i) m(\bar{z}_i)^{-1} \left[\sum_{s=1}^{q_1} h_s^2 B_{1s,i} + \sum_{s=1}^{r_1} \lambda_s B_{2s,i} \right] \bar{M}(\bar{z}) \right\}. \\
&= \int \left\{ \left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} \lambda_s B_{2s}(\bar{z}) \right]' m(\bar{z}_i)^{-1} \left[\sum_{s=1}^{q_1} h_s^2 B_{1s}(\bar{z}) + \sum_{s=1}^{r_1} \lambda_s B_{2s}(\bar{z}) \right] \right\} \bar{M}(\bar{z}) d\bar{z}. \qquad \text{(A.9)}
\end{aligned}
$$

Hence, the leading term of CV_1 is given by (A.9). □

Lemma A.2. *Equation* (A.5) *holds true.*

Proof. We use CV_2^0 to denote CV_2 with $\hat{A}(z_i)^{-1}$ being replaced by $m(z_i)^{-1}$. Using Lemma A.5 it is fairly straightforward to show that $CV_2 = CV_2^0 + (s.o.)$. Hence, we only need to consider CV_2^0,

which we write as

$$
\begin{aligned}
CV_2^0 &= n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} X_i'\mu(z_i)^{-1}X_j u_j K_{\gamma,ij} X_i'\mu(z_i)^{-1}X_l u_l K_{\gamma,il} M_i \\
&= n^{-3}\sum_i\sum_{j\neq i}(X_i'\mu(z_i)^{-1}X_j)^2 u_j^2 K_{\gamma,ij}^2 M_i \\
&\quad + n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq j, l\neq j} X_i'\mu(z_i)^{-1}X_j u_j K_{\gamma,ij} X_i'\mu(z_i)^{-1}X_l u_l K_{\gamma,il} M_i \\
&\equiv D_1 + D_2.
\end{aligned}
$$

We first consider D_2 which has mean zero. By noting that terms related to $\tilde{K}_{\tilde{\gamma},ij}$ are bounded since the kernel functions $W(\cdot)$ and $L(\cdot)$ are bounded and that $h_{q_1+1}\dots h_q$ from the $\tilde{K}$ and from $\mu(\cdot)^{-1}$ cancel, then it is fairly straightforward to show that $E[D_2^2] = n^{-6}O(n^4(h_1\dots h_{q_1})^{-1}) = O((n^2 h_1\dots h_{q_1})^{-1})$. Hence, $D_2 = O_p(n^{-1}(h_1\dots h_{q_1})^{-1/2}) = o_p((nh_1\dots h_{q_1})^{-1})$.

Next, we consider D_1. D_1 can be written as a second order U-statistic. By the U-statistic H-decomposition it is fairly straightforward to show that $D_1 = E(D_1) + (s.o.)$. Define $\delta(\bar{z}_i,\bar{z}_j) = E[(x_i'\mu(\bar{z}_i)^{-1}x_j)^2\sigma^2(x_j,\bar{z}_j)|\bar{z}_i,\bar{z}_j]$. By Lemma A.5 we know that $h_s\to 0$ for $s=1,\dots,q_1$ and $\lambda_s\to 0$ for $s=1,\dots,r_1$. Applying the Taylor expansion and using the law of iterated expectations, and recalling that $\mu(z) = m(\bar{z})\nu_1(\tilde{z})$ and $\nu_2(\tilde{z}) = E[\tilde{K}_{ij}^2|\tilde{z}_i=\tilde{z}]$, we have

$$
\begin{aligned}
&E\left[(x_i'\mu(\bar{z}_i)^{-1}x_j)^2\sigma^2(x_j,\bar{z}_j)K_{\gamma,ij}^2|z_i=z\right] = E\left[\delta(\bar{z}_i,\bar{z}_j)\bar{K}_{\tilde{\gamma},ij}^2|\bar{z}_i=\bar{z}\right]\nu_2(\tilde{z})/\nu_1(\tilde{z})^2 \\
&= (nh_1\dots h_{q_1})^{-1}\kappa^{q_1}\delta(\bar{z},\bar{z})\nu_2(\tilde{z})/\nu_1(\tilde{z})^2 + O_p\left(\zeta_n^{1/2}(nh_1\dots h_{q_1})^{-1}\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E(D_1) &= n^{-1}E[(X_i'\mu(\bar{z}_i)^{-1}X_j)^2\sigma^2(x_j,\bar{z}_j)K_{\gamma,ij}^2 M_i] \\
&= (nh_1\dots h_{q_1})^{-1}\kappa^{q_1}\int \bar{f}(\bar{z})\tilde{f}(\tilde{z})\delta(\bar{z},\bar{z})\nu_2(\tilde{z})\nu_1(\tilde{z})^{-2}M(z)dz + O\left(\zeta_n^{1/2}(nh_1\dots h_{q_1})^{-1}\right).
\end{aligned}
$$

This completes the proof of Lemma A.2 by noting that $\tilde{R}(\tilde{z}) = \nu_2(\tilde{z})/\nu_1(\tilde{z})^2$. □

Lemma A.3. *Equation* (A.6) *holds true.*

Proof. Let CV_3^0 denote CV_3 with $\hat{A}(z_i)$ being replaced by $\mu(z_i)$. Then CV_3^0 is the leading term of

CV_3. We have

$$\begin{aligned}
CV_3^0 &= n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} X_i'\mu(z_i)^{-1}X_j(\beta_j-\beta_i)K_{\gamma,ij}X_i'\mu(z_i)^{-1}X_l u_l K_{\gamma,il} \\
&= n^{-3}\sum_i\sum_{j\neq i} X_i'\mu(z_i)^{-1}X_j(\beta_j-\beta_i)X_i'\mu(z_i)^{-1}X_l u_j K_{\gamma,ij}^2 \\
&\quad + n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i,l\neq j} X_i'\mu(z_i)^{-1}X_j(\beta_j-\beta_i)K_{\gamma,ij}X_i'\mu(z_i)^{-1}X_l u_l K_{\gamma,il} \\
&= CV_{3,1}+CV_{3,2}.
\end{aligned}$$

$CV_{3,1}$ has zero mean. Its second moment is $E[CV_{3,1}^2] = n^{-6}O(n^3(h_1\dots h_{q_1})^{-3}) = O((nh_1\dots h_{q_1})^{-3})$. Hence, $CV_{3,1} = O_P((nh_1\dots h_{q_1})^{-3/2}) = o_p((nh_1\dots h_{q_1})^{-1})$.

$C_{3,2}$ also has zero mean and its second moment is $E[C_{3,2}^2] = n^{-6}O(\zeta_n(n^5+n^4(h_1\dots h_{q_1})^{-1})) = O(n^{-1}\zeta_n)$. Hence, $CV_{3,2} = O_p(n^{-1/2}\zeta_n^{1/2}) = o_p(\zeta_n)$ since $n^{-1/2} = o(\zeta_n^{1/2})$ ($\zeta_n = |\bar{h}|^4+|\bar{\lambda}|^2$). □

Lemma A.4. $CV_{0,2} = o_p(\zeta_n + (nh_1\dots h_{q_1})^{-1})$.

Proof. Letting $CV_{0,2}^0$ denote $CV_{0,2}$ with $\mu(z_i)$ replacing $\hat{A}(z_i)$ in $CV_{0,2}$, we get

$$\begin{aligned}
CV_{0,2}^0 &= n^{-1}\sum_i x_i'\mu(z_i)^{-1}\left[\hat{B}_i+\hat{C}_i\right] \\
&= n^{-2}\sum_i\sum_{j\neq i} u_i x_i'\mu(z_i)^{-1}x_j x_j'(\beta_j-\beta_i)K_{\gamma,ij} + n^{-2}\sum_i\sum_{j\neq i} u_i X_i'\mu(z_i)^{-1}x_j u_j K_{\gamma,ij} \quad = F_1+F_2.
\end{aligned}$$

It is fairly straightforward to show that $E[F_1^2] = n^{-4}O(n^3\zeta_n + n^2(h_1\dots h_{q_1})^{-1}\zeta_n^{1/2}) = O(n^{-1}\zeta_n + n^{-1}\zeta_n^{1/2}(nh_1\dots h_{q_1})^{-1})$. Hence, $F_1 = o_p(\zeta_n + (nh_1\dots h_{q_1})^{-1})$ since $n^{-1} = o(\zeta_n + (nh_1\dots h_{q_1})^{-1})$.

Also, F_2 is a second order degenerate U-statistic, so it is fairly straightforward to show that $E(F_2^2) = n^{-4}O(n^2(h_1\dots h_{q_1})^{-1}) = O((n^2h_1\dots h_{q_1})^{-1})$. Hence, $F_1 = O_p((n^2h_1\dots h_{q_1})^{-1/2}) = o_p((nh_1\dots h_{q_1})^{-1})$. □

Lemma A.5. *Defining* $m(\bar{z}) = E[X_jX_j'|\bar{z}_j = \bar{z}]$, $\nu_1(\tilde{z}) = E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i = \tilde{z}]$ *and* $\mu(z) = m(\bar{z})\nu_1(\tilde{z})$, *then*

$$\hat{A}(z)^{-1} = \mu(z)^{-1} - \mu(z)^{-1}\left[\hat{A}(z)-\mu(z)\right]\mu(z)^{-1} + O_p\left(|\bar{h}|^4+|\bar{\lambda}|^2+\ln(n)(nh_1\dots h_{q_1})^{-1}\right),$$

uniformly in z.

Proof. Defining $\hat{\mu}(z_i) = E[\hat{A}(z_i)|z_i = z]$, then by the independence of $\tilde{z}_i$ and $(y_i, x_i, \bar{z}_i)$, we have that

$$\begin{aligned}\hat{\mu}(z) &= E\left[X_j X_j' \bar{K}_{\bar{\gamma},ij} | \bar{z}_j = \bar{z}\right] E\left[\tilde{K}_{\tilde{\gamma},ij} | \tilde{z}_i = \tilde{z}\right] \\ &= \left\{E[X_j X_j' | \bar{z}_j = \bar{z}] + O\left(|\bar{h}|^2 + |\bar{\lambda}|\right)\right\} E\left[\tilde{K}_{\tilde{\gamma},ij} | \tilde{z}_i = \tilde{z}\right] = \mu(z_i) + O_p\left(|\bar{h}|^2 + |\bar{\lambda}|\right). \qquad \text{(A.10)}\end{aligned}$$

Also, $\hat{A}(z) - \hat{\mu}(z)$ has zero mean and following the standard arguments of deriving uniform convergence rate of nonparametric kernel estimators (e.g., Masry (1996)), we know that

$$\hat{A}(z) - \hat{\mu}(z) = O_p\left(\frac{(\ln(n))^{1/2}}{(nh_1 \dots h_{q_1})^{1/2}}\right) \qquad \text{(A.11)}$$

uniformly in z.

Combining (A.10) and (A.11) we get

$$\hat{A}(z) - \mu(z) = O_p\left(|\bar{h}|^2 + |\bar{\lambda}| + (\ln(n))^{1/2}(nh_1 \dots h_{q_1})^{-1/2}\right). \qquad \text{(A.12)}$$

uniformly in z.

Using (A.12) we obtain

$$\begin{aligned}\hat{A}(z)^{-1} &= \left[\mu(z) + \hat{A}(z) - \mu(z)\right]^{-1} \\ &= \mu(z)^{-1} - \mu(z)^{-1}\left[\hat{A}(z) - \mu(z)\right]\mu(z)^{-1} + O_p\left(|\hat{A}(z) - \mu(z))|^2\right) \\ &= \mu(z)^{-1} - \mu(z)^{-1}\left[\hat{A}(z) - \mu(z)\right]\mu(z)^{-1} + O_p\left(|\bar{h}|^4 + |\bar{\lambda}|^2 + \ln(n)(nh_1 \dots h_{q_1})^{-1}\right),\end{aligned}$$

completing the proof of Lemma A.5. □

Lemma A.6. *$\hat{h}_s = o_p(1)$ for $s = 1, \dots, q_1$ and $\lambda_s = o_p(1)$ for $s = 1, \dots, r_1$.*

Proof. Without assuming any of the smoothing parameters converge to zero, then the only possible non $o_p(1)$ term of $CV(\gamma)$ is a CV_1. It is fairly straightforward to see that $CV_1 = \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i} X_i'\hat{\mu}(z_i)^{-1}X_j X_j'(\beta_j - \beta_i)K_{\gamma,ij}X_i'\hat{\mu}(z_i)^{-1}X_l X_l'(\beta_l - \beta_i)K_{\gamma,il}M_i + o_p(1) \equiv G_1 + o_p(1)$. Note that G_1 can be written as a third order U-statistic and by the H-decomposition of a U-statistic, it is fairly straightforward to show that $G_1 = E(G_1) + o_p(1)$. Further by the law of iterated expectations we have

$$\begin{aligned}E(G_1) &= E\left\{\left[X_i'\hat{\mu}(z_i)^{-1}E\left(X_j X_j'(\beta_j - \beta_i)K_{\gamma,ij}|X_i, Z_i\right)\right]^2 M(Z_i)\right\} \\ &= \int\left[x'(\mu_\beta(\bar{z}) - \beta(\bar{z}))\right]^2 f(x, \bar{z})\bar{M}(\bar{z})d\bar{z}dx \qquad \text{(A.13)}\end{aligned}$$

$\mu_\beta(\bar{z})$ is defined above (15), and $\bar{M}(\bar{z})$ is defined in (21). Note that the right hand side of (A.13)

does not depend on $(h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r)$ since $E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i]$ of the numerator cancels the same quantity from denominator (from $\hat{\mu}(z_i)^{-1} = E[X_j X_j' \bar{K}_{\bar{\gamma},ij}|\bar{z}_i]^{-1} E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{z}_i]^{-1}$).

If the smoothing parameters $h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{r_1}$ that minimize $CV(\gamma)$ do not all converge in probability to zero, then by (15), $E(G_1)$ (or CV_1) does not converge to zero, which implies that the probability that minimum of CV_1, over the smoothing parameters, exceeds δ, does not converge to zero as $n \to \infty$ (for some $\delta > 0$).

However, choosing $h_1, \ldots, h_{q_1}$ to be the size of $n^{-1/(q_1+4)}$, and $\lambda_1, \ldots, \lambda_{r_1}$ to be the size $n^{-2/(q_1+4)}$, letting $h_{q_1+1}, \ldots, h_q$ diverge to infinity, and letting $\lambda_{r_1+1}, \ldots, \lambda_r$ converge to 1, one can easily show that CV_1 converge in probability to zero. This contradicts the result obtained in the previous paragraph, and thus demonstrate that, at the minimum of $CV(\gamma)$, the smoothing parameters $h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{r_1}$, for the relevant components of Z, all converge in probability to zero. □