



Centre for Efficiency and Productivity Analysis

**Working Paper Series
No. WP14/2010**

Local Maximum Likelihood Techniques with Categorical Data

Byeong U. Park, L'opold Simar & Valentin Zelenyuk

Date: December 2010

**School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia**

ISSN No. 1932 - 4398

Local Maximum Likelihood Techniques with Categorical Data

Byeong U. Park*

Department of Statistics
Seoul National University, Korea

Léopold Simar[§]

Institut de Statistique
Université Catholique de Louvain, Belgium

Valentin Zelenyuk

Centre for Efficiency and Productivity Analysis and School of Economics
University of Queensland, Australia.

December 22, 2010

Abstract

In this paper we provide asymptotic theory of local maximum likelihood techniques for estimating a regression model where some regressors are discrete. Our methodology and theory are particularly useful for models that give us a likelihood of the unknown functions we can use to identify and estimate the underlying model. This is the case when the conditional density of the variable of interest, given the explanatory variables, is known up to a set of unknown functions. Examples of such models include probit and logit models, truncated regression models, stochastic frontier models, etc. In developing the theory we use the Racine and Li (2004) kernels for discrete regressors. The asymptotic properties of the resulting estimator are derived and the method is illustrated in various simulated scenarios. The results indicate a great flexibility of the approach and good performances in various complex scenarios, even with moderate sample sizes.

Key words : Local Maximum Likelihood, Nonparametric smoothing, Categorical variables, Truncated regression, Stochastic frontier models.

JEL : Classification: C13, C14, C2

*Research of B. U. Park was supported by the NRF Grant funded by the Korea government (MEST) (No. 20100017437)

[§]L. Simar acknowledges support from the “Interuniversity Attraction Pole”, Phase VI (No. P6/03) of the Belgian Science Policy and from the INRA-GREMAQ, Toulouse, France.

1 Introduction

Recent works have extended the use of nonparametric techniques based on kernel methods to the case where the kernels are applied to categorical and/or ordered discrete variables. This has been done, by using kernel scheme inspired by Aitchison and Aitken (1976), for regression analysis, density estimation, estimation of conditional densities, etc. The local least squares theory is now well developed (e.g., see Racine and Li, 2004 and Li and Racine, 2007), but to the best of our knowledge, analogous theory for the local maximum likelihood approach has not been presented so far.

There are however many econometric models where the information contained in the likelihood function is useful for identifying and estimating the model. Just to give two examples we can consider:

- (i) *Nonparametric stochastic frontier models*: Consider for instance the following production model where Y is the logarithm of the output and $\mathbf{X}$ a vector of logarithm of the inputs

$$Y = r(\mathbf{X}) - u + v, \quad (1.1)$$

where $u|\mathbf{X} = \mathbf{x} \sim |N(0, \sigma_u^2(\mathbf{x}))|$ and $v|\mathbf{X} = \mathbf{x} \sim N(0, \sigma_v^2(\mathbf{x}))$. This decomposition of the error term in a two-sided (v) and a one-sided (u) random variable is important to identify noise from the inefficiency. The functions $r(\mathbf{x})$, $\sigma_u^2(\mathbf{x})$ and $\sigma_v^2(\mathbf{x})$ are unknown. The conditional pdf of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$\text{pdf}(y|\mathbf{x}) = \frac{2}{\sigma(\mathbf{x})} \varphi\left(\frac{y - r(\mathbf{x})}{\sigma(\mathbf{x})}\right) \Phi\left(-(y - r(\mathbf{x})) \frac{\lambda(\mathbf{x})}{\sigma(\mathbf{x})}\right)$$

where $\sigma^2(\mathbf{x}) = \sigma_u^2(\mathbf{x}) + \sigma_v^2(\mathbf{x})$ and $\lambda(\mathbf{x}) = \sigma_u(\mathbf{x})/\sigma_v(\mathbf{x})$.

- (ii) *Nonparametric truncated regressions*: Consider now the model

$$Y = r(\mathbf{X}) + \varepsilon \geq b,$$

for some unknown function r and a known positive constant b . In this model, ε , conditionally on $\mathbf{X} = \mathbf{x}$, has a known continuous distribution $G(\cdot, \tau(\mathbf{x}))$ truncated below at $b - r(\mathbf{x})$, where $\tau(\mathbf{x})$ is unknown. Here, the conditional density of Y given $\mathbf{X} = \mathbf{x}$ equals

$$\text{pdf}(y|\mathbf{x}) = \frac{g_\varepsilon(y - r(\mathbf{x}), \tau(\mathbf{x}))}{1 - G(b - r(\mathbf{x}), \tau(\mathbf{x}))} \mathbb{I}(y \geq b),$$

where $g_\varepsilon(\varepsilon, \tau(\mathbf{x})) = \partial G(\varepsilon, \tau(\mathbf{x}))/\partial \varepsilon$.

We could also mention Generalized Linear Models where specific approaches have already been developed (see e.g; Fan et al., 1995). In all these models, the basic tool for estimating

the functional parameters is that the conditional pdf of Y given $\mathbf{X} = \mathbf{x}$ that can be written as $\text{pdf}(y|\mathbf{x}) = g(y, \boldsymbol{\theta}(\mathbf{x}))$, where $g(\cdot)$ is a known function and $\boldsymbol{\theta}(\mathbf{x}) \in \mathbb{R}^k$ is a set of k unknown functions of $\mathbf{x}$ (this class of models has also been investigated by Severini and Wong, 1992 in the case where a part of the functional parameters is constant). The aim of local maximum likelihood methods is to use local polynomial approximations of these functionals and then to maximize locally the likelihood function. So far these methods have been applied to the case where the explanatory variables $\mathbf{X}$ are of continuous type (see for stochastic frontiers, Kumbhakar et al., 2007 and for truncated regression, Park et al., 2008). Frölich (2006) investigates local likelihood methods in the particular case of a Logit regression model where discrete explanatory variables are allowed, but the statistical properties of the resulting estimators are not derived. Also, due to the binary nature of the dependent variable, the variance function is determined by the regression function.

The aim of this paper is to analyze in the most general setup how local maximum likelihood methods can be adapted when we have continuous $\mathbf{X}$ complemented by a vector of categorical or unordered discrete variables $\mathbf{Z}$. The case of ordered discrete variables will not be developed here, but may be worked out similarly using an appropriate discrete kernel such as in Racine and Li (2004).

So, the conditional pdf of Y will be written as $\text{pdf}(y|\mathbf{x}, \mathbf{z}) = g(y, \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}))$, with $\boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^k$. It should be noticed that here, the k functions in $\boldsymbol{\theta}(\mathbf{x}, \mathbf{z})$ not only cover the regression function, but also all the others parts of the model, including the variance functions. The main contribution of the paper is to provide the asymptotic properties of the resulting estimator in this general setup. We also comment on how to implement the estimator and we illustrate how it works in some simulated examples

The paper is organized as follows. Section 2 presents the methodology, whereas our main results (asymptotic normality) is given in Section 3. Section 4 illustrates by some simulated data how the estimator behaves in practice in finite samples situations, including the selection of the bandwidths. The technical details, regularity conditions and proofs are in Section 5. Section 6 concludes.

2 Methodology

Suppose we observe $(\mathbf{X}^i, \mathbf{Z}^i, Y^i)$, $1 \leq i \leq n$, which are i.i.d. copies of $(\mathbf{X}, \mathbf{Z}, Y)$, where $\mathbf{X} \in \mathbb{R}^p$ is a vector of p continuous random variables, $\mathbf{Z} \in \mathbb{R}^d$ is a vector of d discrete random variables, where the domain of Z_j is $\{0, 1, \dots, c_j - 1\}$, and Y is a random variable of any type. Let $g(\cdot, \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}))$ for a function $\boldsymbol{\theta}$, whose values are in $\mathbb{R}^k$, denote the conditional density function of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. Suppose that the functional form of $g(\cdot, \cdot)$ is known.

We are interested in estimating the vector of k multivariate functions $\boldsymbol{\theta}$.

The log-likelihood function of $\boldsymbol{\theta}$ is then given by

$$L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log g(Y^i, \boldsymbol{\theta}(\mathbf{X}^i, \mathbf{Z}^i)).$$

Direct maximization of the likelihood over $\boldsymbol{\theta}$ in an infinite-dimensional function space is intractable, and suffers from overfitting. The main idea of our method is to approximate $\boldsymbol{\theta}(\mathbf{u}, \mathbf{v})$ in a neighborhood of $(\mathbf{x}, \mathbf{z})$ by a local polynomial in the direction of $\mathbf{x}$, and then maximize the resulting local likelihood function at the point $(\mathbf{x}, \mathbf{z})$. In this paper we focus on the local linear case. Extension to the general order of local polynomial fitting requires additional notational complexity, but the main idea is the same. For the local linear fitting, we take the following approximation: $\boldsymbol{\theta}(\mathbf{u}, \mathbf{v}) \simeq \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z})(\mathbf{u} - \mathbf{x})$, where $\boldsymbol{\Theta}(\mathbf{x}, \mathbf{z})$ is a $(k \times p)$ -dimensional matrix. To take only neighboring observations around $(\mathbf{x}, \mathbf{z})$, or to give more weights on them, in the construction of a local likelihood, we employ a kernel approach. Define $K_{\mathbf{H}} = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1} \cdot)$ for a nonnegative function K , called kernel, defined on $\mathbb{R}^p$ and a symmetric positive definite matrix $\mathbf{H}$, called bandwidth matrix, where $|\mathbf{H}|$ denotes the determinant of $\mathbf{H}$. Also, define

$$\Lambda_{\mathbf{w}}(\mathbf{z}', \mathbf{z}) = \prod_{j=1}^d w_j^{I(z'_j \neq z_j)},$$

for a d -vector $\mathbf{w} = (w_j)$ with $0 \leq w_j \leq 1$, where $I(A)$ is an indicator such that $I(A) = 1$ if A holds, and 0 otherwise. The local (linear) likelihood at $(\mathbf{x}, \mathbf{z})$ is then given by

$$L_n(\boldsymbol{\theta}, \boldsymbol{\Theta}; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^n \log g(Y^i, \boldsymbol{\theta} + \boldsymbol{\Theta}(\mathbf{X}^i - \mathbf{x})) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}). \quad (2.1)$$

Let $(\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}), \hat{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{z}))$ maximize the local likelihood $L_n(\boldsymbol{\theta}, \boldsymbol{\Theta}; \mathbf{x}, \mathbf{z})$. The proposed estimator of $\boldsymbol{\theta}(\mathbf{x}, \mathbf{z})$ is then $\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$, while $\hat{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{z})$ gives an estimator of the 1st partial derivative of $\boldsymbol{\theta}$ with respect to $\mathbf{x}$. The above kernel scheme for the discrete covariates $\mathbf{Z}^i$ has been also employed by Racine and Li (2004), inspired by Aitchison and Aitken (1976). We note that a slightly different kernel defined by $\Lambda_{\mathbf{w}}(\mathbf{z}', \mathbf{z}) = \prod_{j=1}^d (w_j / (c_j - 1))^{I(z'_j \neq z_j)} (1 - w_j)^{I(z'_j = z_j)}$ for $0 \leq w_j \leq 1 - 1/c_j$ is equivalent to ours. This is because

$$\prod_{j=1}^d \left(\frac{w_j}{c_j - 1} \right)^{I(z'_j \neq z_j)} (1 - w_j)^{I(z'_j = z_j)} = \left[\prod_{j=1}^d (1 - w_j) \right] \times \prod_{j=1}^d \left(\frac{w_j}{(c_j - 1)(1 - w_j)} \right)^{I(z'_j \neq z_j)}$$

so that $w_j / ((c_j - 1)(1 - w_j))$ takes the role of w_j in the definition of our kernel.

We also note that approximation of $\boldsymbol{\theta}$ in the direction of $\mathbf{z}$ does not make sense since $\mathbf{Z}$ takes values whose distances are fixed. Basically, we are approximating $\boldsymbol{\theta}(\mathbf{X}^i, \mathbf{Z}^i)$ by

$\boldsymbol{\theta}(\mathbf{x}, \mathbf{Z}^i) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{Z}^i)(\mathbf{X}^i - \mathbf{x})$. This consideration may lead to the following localized log-likelihood:

$$\begin{aligned} & \sum_{i=1}^n \log g(Y^i, \boldsymbol{\theta}(\mathbf{x}, \mathbf{Z}^i) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{Z}^i)(\mathbf{X}^i - \mathbf{x})) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) I(\mathbf{Z}^i = \mathbf{z}) \\ &= \sum_{i=1}^n \log g(Y^i, \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z})(\mathbf{X}^i - \mathbf{x})) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) I(\mathbf{Z}^i = \mathbf{z}), \end{aligned}$$

which coincides with the one at (2.1) when all $w_j = 0$ in which case $\Lambda_{\mathbf{w}}(\mathbf{z}', \mathbf{z}) = I(\mathbf{z}' = \mathbf{z})$. (Here, we adopt the convention $0^0 = 1$.) The main drawback of the latter is that, for each $\mathbf{z}$, there may not be enough observations such that $\mathbf{Z}^i = \mathbf{z}$ to conduct nonparametric estimation in the direction of $\mathbf{x}$. This is particularly the case when some components of $\mathbf{Z}$ assume a large number of discrete values. Even in the case where every component of $\mathbf{Z}$ takes a small number of discrete values, there may not be enough observations such that $\mathbf{Z}^i = \mathbf{z}$, especially when the dimension of $\mathbf{Z}$ is high. Taking $w_j > 0$ in (2.1) avoids this difficulty. For consistency we need $w_j \rightarrow 0$, however, as the sample size n goes to infinity. We will be more specific for the rate of convergence in the next section. Finally, we remark that another extreme choice $w_j = 1$ leads to ‘no localization’ in the direction of $\mathbf{z}$ and thus ignorance of the presence of $\mathbf{Z}$, since it gives $\Lambda_{\mathbf{w}}(\mathbf{z}', \mathbf{z}) = 1$ for all $\mathbf{z}, \mathbf{z}'$.

3 Theoretical Properties

In this section we give the asymptotic distribution of $\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. For this we introduce some notations. Let f denote the density function of $(\mathbf{X}, \mathbf{Z})$. Let $d(\mathbf{z}', \mathbf{z}) = \sum_{j=1}^d I(z'_j \neq z_j)$. Define $\mathbf{g}_1(y, \boldsymbol{\theta}) = \partial \log g(y, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, which is a k -dimensional vector of functions. Also, let $\mathbf{g}_2(y, \boldsymbol{\theta})$ denote the Hessian matrix of $\log g(y, \boldsymbol{\theta})$, i.e., $\mathbf{g}_2(y, \boldsymbol{\theta}) = \partial^2 \log g(y, \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$. Define

$$\begin{aligned} \boldsymbol{\rho}(\mathbf{x}, \mathbf{z}) &= -E[\mathbf{g}_2(Y, \boldsymbol{\theta}(\mathbf{X}, \mathbf{Z})) | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}], \\ \boldsymbol{\tau}(\mathbf{x}, \mathbf{z}) &= E[\mathbf{g}_1(Y, \boldsymbol{\theta}(\mathbf{X}, \mathbf{Z})) \mathbf{g}_1(Y, \boldsymbol{\theta}(\mathbf{X}, \mathbf{Z}))^\top | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}]. \end{aligned}$$

Let $\boldsymbol{\theta}_j''$ denote the Hessian matrix of the function θ_j , the j th component of the true $\boldsymbol{\theta}$. Let $\mathcal{D}$ and $\mathcal{D}_j$, respectively, denote the sets of vectors and real numbers that $\mathbf{Z}$ and Z_j can assume. We consider a spherically symmetric (around zero) multivariate kernel K such that

$$\int K(\mathbf{u}) d\mathbf{u} = 1, \quad \int \mathbf{u} \mathbf{u}^\top K(\mathbf{u}) d\mathbf{u} = \mu_2 \mathbf{I}_p$$

for some positive constant μ_2 , where $\mathbf{I}_p$ denotes the identity matrix of dimension p . The first property is satisfied by any nonnegative kernels after normalization. The second one we assume for simplicity of presentation. It is satisfied by a product kernel which is the product

of symmetric univariate kernels. We also assume that K is supported on a compact set with nonempty interior in $\mathbb{R}^p$.

For the bandwidth matrix $\mathbf{H}$, we assume that all entries tends to zero and $n|\mathbf{H}|$ grows to infinity as n goes to infinity. Also, we assume that $(n|\mathbf{H}|)^{1/2}\text{tr}(\mathbf{H}^2)$ is bounded. Note that $\text{tr}(\mathbf{H}^2)$ is simply the squared Frobenius norm of the matrix $\mathbf{H}$. In the case where $\mathbf{H} = h\mathbf{I}_p$ for a scalar bandwidth h , the condition means $h = O(n^{-1/(p+4)})$. For the weights of the discrete kernel, we assume that each w_j tends to zero as n goes to infinity. Also, we assume $w^* := \max_{1 \leq j \leq d} w_j = O(\text{tr}(\mathbf{H}^2))$. The following theorem is for fixed points $\mathbf{x}$ and $\mathbf{z}$. In the theorem, $\mathbf{z}_{-j}$ is the $(d-1)$ -vector which is obtained by deleting the j th entry of $\mathbf{z}$.

Theorem 3.1. *Assume the conditions in Section 5.1. Then, we have*

$$(n|\mathbf{H}|)^{1/2} \left[\frac{1}{f(\mathbf{x}, \mathbf{z})} \int K^2(\mathbf{u}) d\mathbf{u} \cdot \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})^{-1} \boldsymbol{\tau}(\mathbf{x}, \mathbf{z}) \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})^{-1} \right]^{1/2} \\ \times \left[\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) - \frac{1}{2} \mu_2 \boldsymbol{\beta}_{\mathbf{H}}(\mathbf{x}, \mathbf{z}) - \mathbf{b}_{\mathbf{w}}(\mathbf{x}, \mathbf{z}) \right] \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k),$$

where $\boldsymbol{\beta}_{\mathbf{H}}(\mathbf{x}, \mathbf{z}) = [\text{tr}(\boldsymbol{\theta}_1''(\mathbf{x}, \mathbf{z})\mathbf{H}^2), \dots, \text{tr}(\boldsymbol{\theta}_k''(\mathbf{x}, \mathbf{z})\mathbf{H}^2)]^\top$ and

$$\mathbf{b}_{\mathbf{w}}(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z})^{-1} \sum_{j=1}^d w_j \sum_{z'_j \neq z_j, z'_j \in \mathcal{D}_j} f(\mathbf{x}, \mathbf{z}_{-j}, z'_j) \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})^{-1} \boldsymbol{\rho}(\mathbf{x}, \mathbf{z}_{-j}, z'_j) [\boldsymbol{\theta}(\mathbf{x}, \mathbf{z}_{-j}, z'_j) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z})].$$

The theorem tells that the first-order properties of the estimator depend on the weights of the discrete kernel $\Lambda_{\mathbf{w}}$ only through the bias. We point out this is due to $w^* = o(1)$, however. The theorem also implies that, if $w^* = o(\text{tr}(\mathbf{H}^2))$, then the leading bias $\mathbf{b}_{\mathbf{w}}(\mathbf{x}, \mathbf{z})$ involving w_j is also negligible. We also note that, if $\mathbf{H} = h\mathbf{I}_p$ for a scalar h , then the j th component of $\boldsymbol{\beta}_{\mathbf{H}}(\mathbf{x}, \mathbf{z})$ equals $h^2 \sum_{l=1}^p \partial^2 \theta_j(\mathbf{x}, \mathbf{z}) / \partial x_l^2$, where x_l is the l th component of $\mathbf{x}$. The latter indicates that the bias increases with the curvature of the θ_j .

It should be noticed that one may derive an analogue of Theorem 3.1 for the derivative estimator $\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. Extending the technical arguments in the proof of Theorem 3.1, one may show that the asymptotic bias of $\hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$ is $O(h^2 + h^{-1}w^*)$ and the asymptotic variance equals $O(n^{-1}h^{-3})$, in the case where $\mathbf{H} = h\mathbf{I}_p$, under sufficient smoothness properties of $\boldsymbol{\theta}$.

The analysis of the effect of the categorical variable could be done according to the following ideas. To simplify the discussion, suppose that the dimension of $\mathbf{Z}$ equals 1 and c_1 in the discrete kernel equals 2. Then, one may be interested in estimating $\boldsymbol{\theta}(\mathbf{x}, z=0) - \boldsymbol{\theta}(\mathbf{x}, z=1)$ by $\hat{\boldsymbol{\theta}}(\mathbf{x}, z=0) - \hat{\boldsymbol{\theta}}(\mathbf{x}, z=1)$ that follows a normal distribution asymptotically. To avoid an explicit calculation of the asymptotic covariance between $\hat{\boldsymbol{\theta}}(\mathbf{x}, z=0)$ and $\hat{\boldsymbol{\theta}}(\mathbf{x}, z=1)$, a bootstrap approach would be helpful.

We remark that, analogous to the case of local least-squares estimators, the rate of convergence does not depend on the presence of the discrete variables (no curse of dimensionality for $\mathbf{Z}$). It should be also noticed that the theorem coincides with the classical result when there are no discrete variables but only continuous ones. Note also that in Park et al. (2008) it was shown, in the truncated regression model, that using lower order polynomial for the variance function deteriorates the rate of convergence of the regression function and that, in the case of constant variance function (a model in the spirit of Severini and Wong, 1992), Park et al. (2008) suggested a 2-stage procedure giving a $\sqrt{n}$ -consistent estimator of the variance. These two results remain valid in the model here.

In practice, the choice of the bandwidths could be done by using a cross-validation (CV) leave-one-out likelihood criterion. We may select the bandwidths $\mathbf{H}$ and $\mathbf{w}$ that maximizes

$$CV(\mathbf{H}, \mathbf{w}) = (1/n) \sum_{i=1}^n \log g(Y^i, \hat{\boldsymbol{\theta}}_{\mathbf{H}, \mathbf{w}}^{(-i)}(\mathbf{X}^i, \mathbf{Z}^i)),$$

where $\hat{\boldsymbol{\theta}}_{\mathbf{H}, \mathbf{w}}^{(-i)}(\mathbf{X}^i, \mathbf{Z}^i)$ is the estimate of the function $\boldsymbol{\theta}$ at the point $(\mathbf{X}^i, \mathbf{Z}^i)$ obtained from the “leave-the i th observation-out” sample of size $(n - 1)$ with the value $(\mathbf{H}, \mathbf{w})$ for the bandwidths. If n is too large, we may select a random subsample of size $m \ll n$ for the evaluation points $(\mathbf{X}^i, \mathbf{Z}^i)$. It is the approach which has been adopted in the section for numerical examples. As in Racine and Li (2004), it remains true that, if there are no continuous variables, the CV choice of $\mathbf{w}$ will converge to zero at the rate n^{-1} .

Note that if $\boldsymbol{\theta}(\mathbf{u}, \mathbf{v}) = \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z})(\mathbf{u} - \mathbf{x})$ exactly, i.e. the “working parametric model” is true, then one may get the parametric rate of convergence by letting $h \rightarrow \infty$.

4 Numerical Illustration

To understand the performances of the estimator, we tried various functional forms and we present here a few examples from the typical simulation results.

4.1 Example 1

First, let us consider the truncated regression case, e.g., assuming $y = r(x, z) + \varepsilon$ with

$$r(x, z) = a_1 + a_2(1 - z) + b_1 \sin(\gamma x) + b_2 xz + b_3(1 - z)x^2,$$

where $\varepsilon \sim N(0, \sigma^2(x, z))$, $\sigma(x, z) = \sigma_c \sqrt{3 - x}$, and $\varepsilon \geq 1 - r(x, z)$. Finally, $x \sim U(-2, 2)$ while $z \in \{0, 1\}$ is generated so that $\Pr(z = 1) = 0.5$.

Note that even with this relatively simple DGP, it might be quite challenging to guess correctly about the correct parametric form of the regression function and of the skedastic

function, needed to achieve consistent parametric estimates. However, with the same amount of *a priori* information, yet with no parametric assumption on the regression function, the local likelihood estimator with discrete kernel handling the discrete variables does a good job, as is illustrated below. Note that in the first two examples we present results for the local likelihood estimators when the variance is approximated linearly, while the regression function has quadratic approximation (denoted with LQMLE) or linear approximation (denoted with LLMLE). As with other kernel-based estimators, the choice of the bandwidth is critical, especially for relatively small samples or when the true regression function is highly non-linear. Here, we present only the maximum-likelihood cross-validation (MLCV) method for simultaneously choosing all the bandwidths, as was described in the preceding section.¹ Also note that in most of cases presented, the MLCV optimizations were done when estimating the regression curve using full sample but evaluated at a set of 50 randomly selected observations from this sample.²

The set of panels of Figure 1 below illustrates the case when parameters are set to $a_1 = 1.5$, $a_2 = 0$, $b_1 = 0.5$, $b_2 = 0$, $b_3 = 0$, $\gamma = \pi$ and, $\sigma_c = 0.15$. Note that in this particular example the two groups have identical relationships. The NW panel presents the results for $n = 100$, when the bandwidth for the discrete kernel, w , is chosen to be 0. This is equivalent to separate estimation for each group. We see that the performance of the estimation is fairly good. However, it can be improved substantially when the estimation is performed with an optimal value of the bandwidth for the discrete kernel (e.g., according to MLCV), which is illustrated for the same sample on the NE panel. Furthermore, the two bottom panels present typical estimation results for the same scenario as the corresponding top panels, but when $n = 200$, illustrating the improvements in performance when increasing the sample size.

Also, note that while the results from the NW panel suggest that separate estimation for each group still yields fairly good results, this of course might not be the case when one of the groups has relatively small number of observations, making separate estimations infeasible or very poor. In this case, choosing an optimal value of w , rather than setting it to zero, would be crucial.

¹We also tried the least squares cross-validation (LSCV) method and the results were similar. Interestingly, yet not so surprisingly, our simulations showed that the LSCV appeared to be somewhat more robust for relatively small samples and faster to optimize for large samples, yet the MLCV method gave better fit for relatively large samples.

²Of course, higher precision of bandwidth estimation can be reached by using entire sample but at much higher computing cost. For example, note that MLCV optimizations performed for 100 out of 100 observations in example 1 took about 7 hours on our desktop PC (Intel Core, Duo CPU E8400 with 3GHz and 2GB RAM), while it took less than 1 hour for 50 out of 100 and even out of 500 observations. Note however that for more complicated scenarios doing MLCV for a small sub-sample might be critical: e.g., for the scenario of example 2, MLCV optimizations for 50 out of 100 observations took about 7 hours.

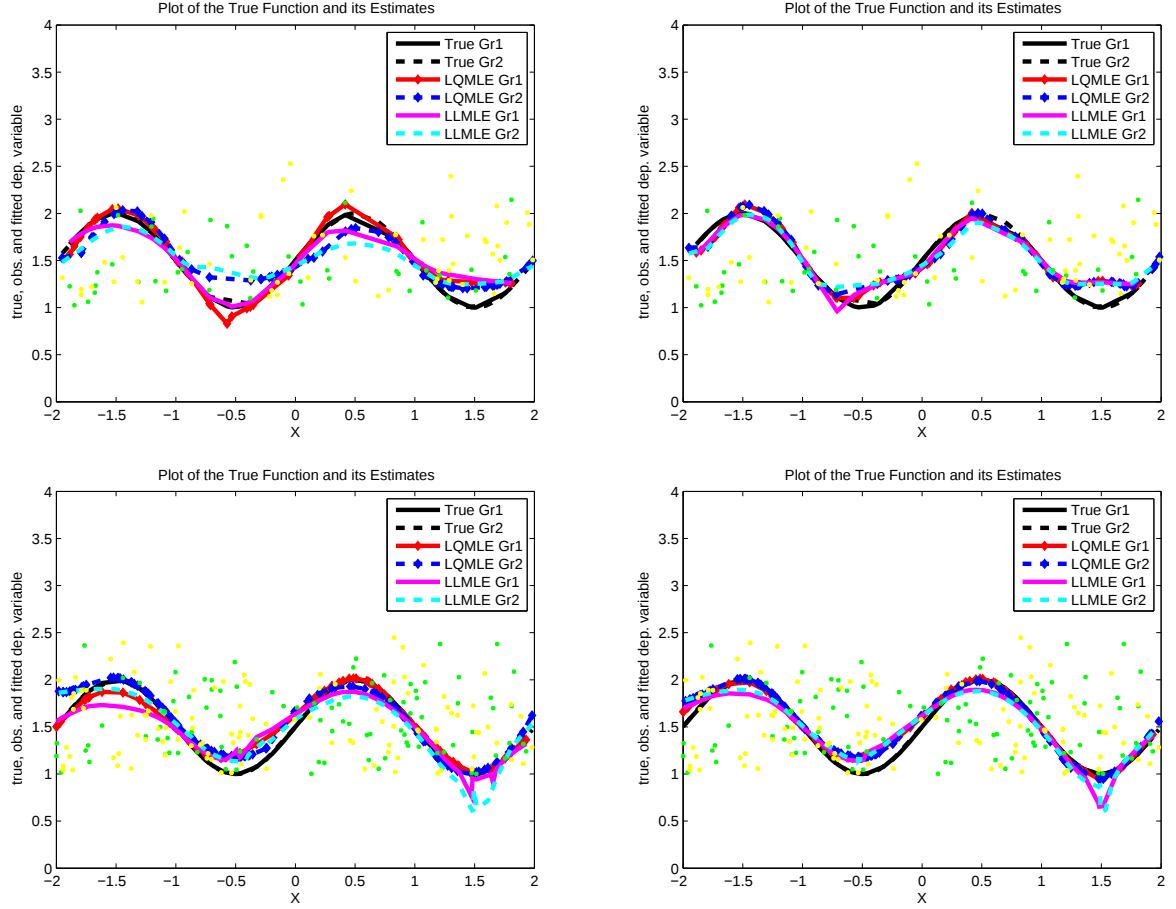


Figure 1: Typical estimation results for scenario of Example 1. NW Panel: separate estimation for each group, $n = 100$; NE panel: estimation with bandwidths chosen via MLCV, $n = 100$; SW Panel: separate estimation for each group, $n = 200$; SE panel: estimation with bandwidths chosen via MLCV, $n = 200$.

4.2 Example 2

In this example, we complicate the previous scenario by making the two groups to have very different relationships, not only by a constant but also by the curvature. Figure 2 below illustrates the cases where the parameters of the scenario described above are set to $a_1 = 1.7$, $a_2 = -0.7$, $b_1 = 0.5$, $b_2 = 0.4$, $b_3 = 0.3$, $\gamma = \pi$ and $\sigma_e = 0.15$. The NW panel shows what happens when the discrete regressor is ignored (or $w = 1$) and $n = 100$, where we see that the estimator is “confused”, trying to find a local average for the two groups together, which is not what the true DGP is about. This panel gives an illustration of the classical problem of omitted variable bias, when the discrete explanatory variable is omitted. The NE panel presents the results for the same sample but with another extreme when $w = 0$, i.e., separate

estimations for each group. The SW panel illustrates the results of the estimation for the same sample but when the bandwidths are chosen via MLCV method. We can see here that the estimator does fairly well, certainly better than estimation under $w = 1$. Comparing the the NE and SW panels suggests that for this particular simulation, the separate estimation ($w = 0$) gave slightly better fit, which is due to the fact that MLCV results attributed very small positive weight (about $w = 0.1$) for sharing information between the two sub-samples. Finally, the SE panel, presents the estimation results when the bandwidths are chosen via MLCV for $n = 200$, to illustarte the substantial improvements with the sample size n increases.

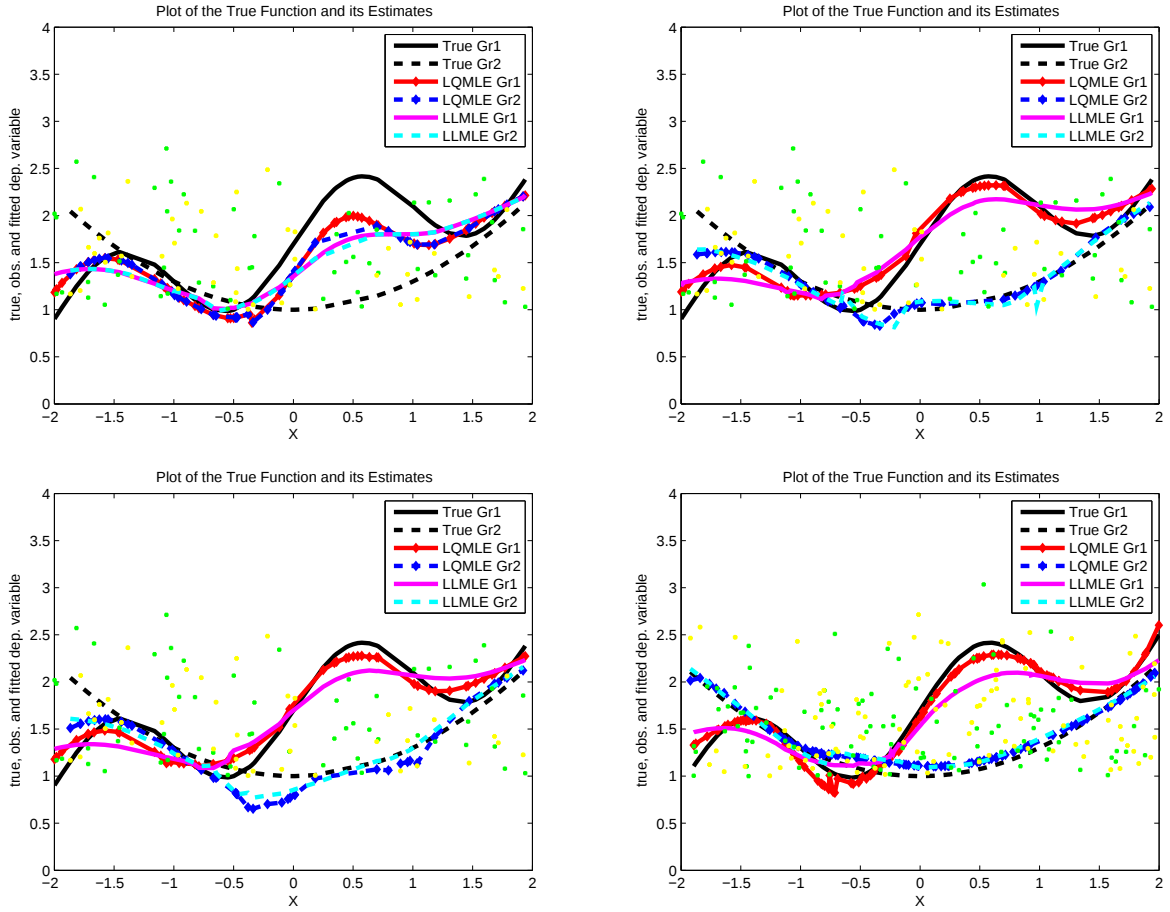


Figure 2: Typical estimation results for scenario of Example 2. NW panel: discrete regressor is ignored (or $w = 1$); NE panel: separate estimation for each group ($w = 0$), $n = 100$; SW panel: estimation with bandwidths chosen via MLCV, $n = 100$; SE panel: estimation with bandwidths chosen via MLCV, $n = 200$.

It is worth noting here that despite the fact that the two groups in the example above have very different curvatures for the regression functions, implying that the optimal level

of smoothing might be different for each group, our estimator with the same h for both groups provides fairly good fit.³ Furthermore, the estimation results were good even in the case (not illustrated here) when one group had a very non-linear regression function while it was linear for the other group (and so the optimal bandwidth for the latter group would be infinite). Finally, note that in most of the examples we had run, including the ones presented here, the local likelihood with quadratic approximation gives sizably better fit than the one with linear approximations, especially for the highly non-linear cases. This conclusion is pertinent to both, the regression function and the variance. This is not surprising, whereas it is often ignored in practice, with applied researchers often just using the local linear (or even constant) approximations in their estimations.

4.3 Example 3

Here we consider the local likelihood estimation of a stochastic frontier analysis model with discrete regressors. Specifically, the econometric model is given by $y = r(x, z) + \varepsilon$ where $\varepsilon = v - u$, $v \sim N(0, \sigma_v^2(x, z))$ and $u \sim |N(0, \sigma_u^2(x, z))|$.

We choose $x \sim U(0.5, 10)$ and $z \in \{0, 1\}$ is so that $\Pr(z = 1) = 0.5$. The reader must recognize a resemblance with the classical Aigner et al. (1977) model, except that the variances of the two terms are allowed to be heteroskedastic. For the purpose of illustration, we assume a non-linear form for the regression relationship, given by

$$r(x, z) = a_1 x + a_2 \sin x + \gamma z.$$

To be coherent with production theory, we set $a_1 = a_2$ to ensure monotonicity of the production frontier, $r(x, z)$, although it is certainly not required for our estimator. We also set $\gamma > 0$ to make group 1 ($z = 1$) having a technology superior than that of group 2. At the same time, we also make the inefficiency term, u , to depend positively on the continuous regressor x for both groups. Intuitively, one could imagine that labor can influence both the maximal output and the level of inefficiency (e.g., due to increased human errors). In addition, we want the group 1 to have also a higher level of inefficiency than that of group 2, *ceteris paribus*. We could interpret this scenario as if one group would consist of firms that never use the technology which is superior but riskier (i.e., having higher probability of failures, reflected in the variance of the inefficiency term). We model such phenomenon via the skedastic function of the inefficiency term given by

$$\sigma_u(x, z) = \sigma_{u0} \exp(\sigma_{u1} \log x + \sigma_{ud} z).$$

³Note, however, that MLCV estimation in such case was very computer intensive and not always robust to initial values, e.g., for 50 out of 200 observations in this particular example it took about 20 hours on our PC.

On the top of all this, we also allow for a heteroskedastic noise,

$$\sigma_v(x, z) = \sigma_{v0} \exp(\sigma_{v1} \log x + \sigma_{vd} z).$$

It is worth noting that it is hardly possible to guess correctly about such scenario to do a parametric estimation, but it also looks quite challenging to estimate it non-parametrically. Indeed, not only the continuous regressor influences both the maximal output and the efficiency level (in opposite direction), but also the discrete regressor determines both the level or choice of technology and the level of efficiency (and also in opposite directions).

We tried many different parameters and the results were qualitatively similar and for the particular set of panels presented in Figure 3, we set $a_1 = a_2 = 1$, $\gamma = 2$, $\sigma_{u0} = 0.3$, $\sigma_{u1} = -0.3$, $\sigma_{ud} = 0.2$ and $\sigma_{v0} = \rho \sigma_{u0}$. Note that ρ reflects the level of noise relative to the level of inefficiency in terms of variances, and, as expected, the higher ρ , the less accurate the fit were. In Figure 3, we present a fair case, when $\rho = \sqrt{(\pi - 2)/\pi}$, and we also set $\sigma_{v1} = \sigma_{u1}$, $\sigma_{vd} = \sigma_{ud}$, which ensures that the noise and the inefficiency contribute equally to the total variance of the unobserved composite error.⁴

In Figure 3, The NW panel illustrates the scenario for $n = 100$ and one can see that the estimator is doing fairly well. Indeed, despite the fact that both the continuous and the discrete regressors are influencing technology and efficiency levels in opposite directions, the estimator is able to capture the unknown structure of the relationships (e.g., the varying returns to scale at different levels of the input, etc.), conveying well the essence of the true model. Furthermore, the NE panel presents estimation for the same scenario for $n = 1000$, and as expected, one can see even better performance. The SW panel presents the estimation for the same scenario, for $n = 100$, but with $\gamma = 0$, i.e., when both groups have the same technology, but different levels (and variances) of the inefficiency term and of the statistical noise. We observe that the estimator captures and presents the essence of the truth quite well even for such a small sample size.⁵ The SE panel presents the same as the latter scenario, but for $n = 1000$ and, as expected, we see substantial improvements in the estimation. Noteworthy, the MLCV optimization yielded $w = 0$ for the top panels, while $w = 0.6509$ and $w = 0.7228$ for the bottom panels (left and right respectively).

⁴Recall that the total conditional variance of ε is given by $\text{Var}(v - u|x, z) = \sigma_v^2(x, z) + \sigma_u^2(x, z)(\pi - 2)/\pi$. Note also that here we used local likelihood estimation with quadratic approximation of $r(x, z)$ and of $\log \sigma_u^2(x, z)$, and linear approximation for $\log \sigma_v^2(x, z)$.

⁵Here, we applied MLCV to the entire sample ($n = 100$) as the computation time was not large for this type of model and the MLCV yielded more precise bandwidths that gave better fit.

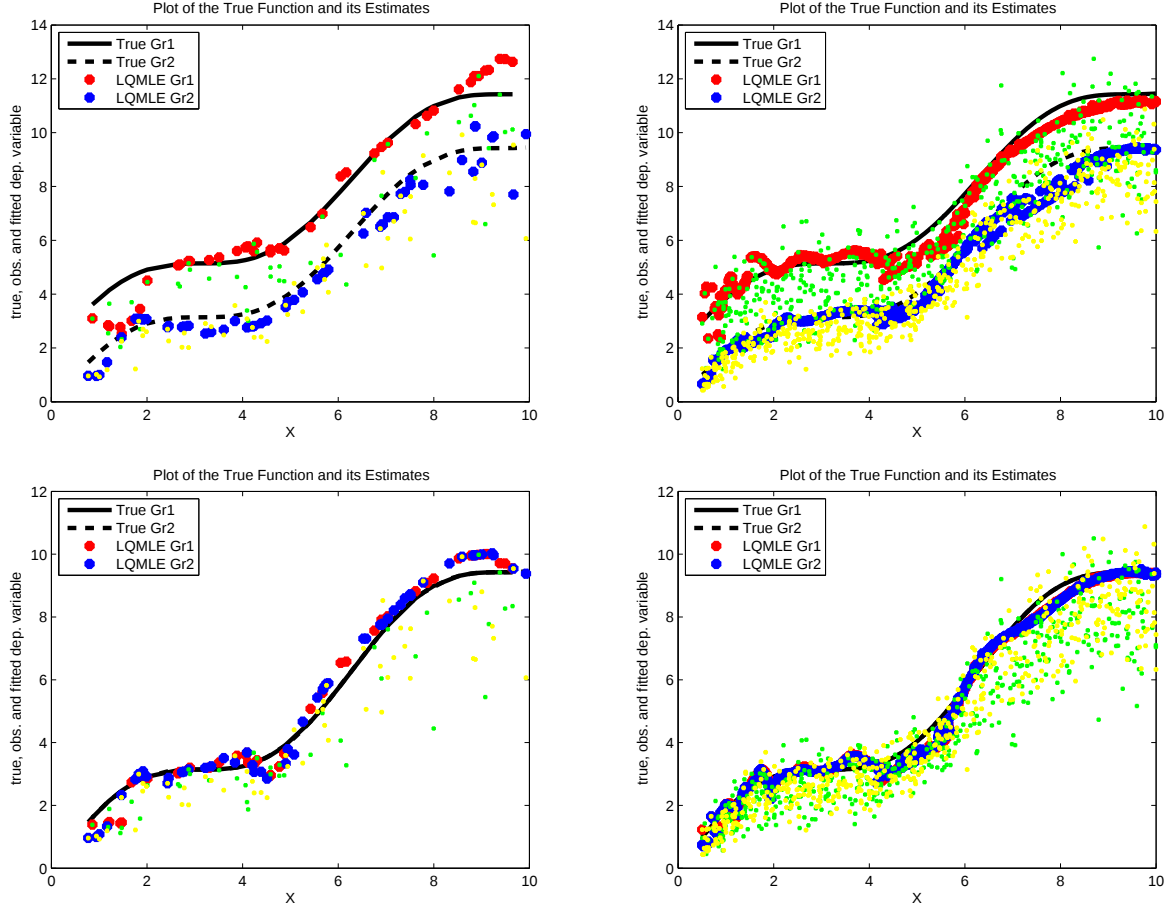


Figure 3: Typical estimation results based on MLCV bandwidths for scenario of Example 3. NW panel: $\gamma = 2$, $n = 100$; NE panel: same but $n=1000$; SW panel: $\gamma = 0$, $n = 100$; SE panel: same but $n = 1000$.

5 Technical Details

Below in the conditions and in the proof of Theorem 3.1, $\|\mathbf{v}\|$ denotes the usual ℓ_2 -norm for a vector $\mathbf{v}$, and the Frobenius (Hilbert-Schmidt) norm for a matrix $\mathbf{v}$. Define $\psi(\mathbf{s}|\mathbf{x}, \mathbf{z}) = E[\mathbf{g}_1(Y, \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \mathbf{s} | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})]$ for $\mathbf{s} \in \mathbb{R}^k$. The conditions and the proof are given for a fixed point $(\mathbf{x}, \mathbf{z})$ at which we want to estimate the value of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$.

5.1 Regularity conditions

(A1) For the vector of functions $\mathbf{G}$ defined at (5.1), the equation $\mathbf{G}(\boldsymbol{\alpha}, \mathbf{A}) = \mathbf{0}$ has the unique solution $\boldsymbol{\alpha} = \mathbf{0}$ and $\mathbf{A} = \mathbf{O}$, where $\mathbf{0}$ is the zero vector and $\mathbf{O}$ is the zero matrix. Also, $E[\mathbf{g}_1(Y, \boldsymbol{\theta}(\mathbf{X}, \mathbf{Z}))|\mathbf{X}, \mathbf{Z}] = \mathbf{0}$ almost surely.

- (A2) For any compact set $\mathcal{C}$, there exists a function U_1 such that $\sup_{\boldsymbol{\theta} \in \mathcal{C}} \|\mathbf{g}_1(y, \boldsymbol{\theta})\| \leq U_1(y)$ and $\sup_{\|\mathbf{u}-\mathbf{x}\| \leq \epsilon} E[U_1(Y)^{2+\delta} | \mathbf{X} = \mathbf{u}] < \infty$ for some $\epsilon, \delta > 0$. Also, $\mathbf{g}_2(y, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for each y , and there exists a function $U_2(y)$ such that $\sup_{\boldsymbol{\theta} \in \mathcal{C}} \|\mathbf{g}_2(y, \boldsymbol{\theta})\| \leq U_2(y)$ for any compact set $\mathcal{C}$ and $\sup_{\|\mathbf{u}-\mathbf{x}\| \leq \epsilon} E[U_2(Y)^2 | \mathbf{X} = \mathbf{u}] < \infty$ for some $\epsilon > 0$.
- (A3) All entries of $\boldsymbol{\theta}(\cdot, \mathbf{v})$ are twice partially continuously differentiable at $\mathbf{x}$ for all values of $\mathbf{v}$ such that $d(\mathbf{v}, \mathbf{z}) = 0$ or 1. Also, there exists $\epsilon > 0$ such that for all $1 \leq j \leq k$
- $$\sup_{\|\mathbf{u}-\mathbf{x}\| \leq \epsilon, \mathbf{v} \in \mathcal{D}} \left\| \frac{\partial}{\partial \mathbf{u}} \theta_j(\mathbf{u}, \mathbf{v}) \right\| < \infty.$$
- (A4) All entries of $\boldsymbol{\rho}(\cdot, \mathbf{v})$ are continuous at $\mathbf{x}$ for all values $\mathbf{v}$ such that $d(\mathbf{v}, \mathbf{z}) = 0$ or 1, and $\boldsymbol{\rho}(\mathbf{x}, \mathbf{z})$ is positive definite.
- (A5) The density function $f(\cdot, \mathbf{v})$ is continuous at $\mathbf{x}$ for all values $\mathbf{v}$ such that $d(\mathbf{v}, \mathbf{z}) = 0$ or 1, and $f(\mathbf{x}, \mathbf{z}) > 0$.
- (A6) All entries of $\boldsymbol{\tau}(\cdot, \mathbf{z})$ is continuous at $\mathbf{x}$.
- (A7) For any compact set $\mathcal{C}$, it holds that $\sup_{\mathbf{s} \in \mathcal{C}} \|\boldsymbol{\psi}(\mathbf{s} | \mathbf{x} + \mathbf{u}, \mathbf{z}) - \boldsymbol{\psi}(\mathbf{s} | \mathbf{x}, \mathbf{z})\| \rightarrow 0$ as $\|\mathbf{u}\| \rightarrow 0$.

The first part of the assumption (A1) is required for likelihood-based methods. Without this assumption, likelihood-based methods would not work. It holds if the logarithm of the conditional density $\log g(y, \boldsymbol{\theta})$ is strictly convex in $\boldsymbol{\theta}$, the latter being typically assumed for likelihood-based methods. The second part of (A1) is also typical. It is just a Bartlett identity of first-order. The two conditions of (A2) are for a stochastic expansion and the asymptotic normality of the estimator. For the stochastic expansion we actually need the first condition with $\delta = 0$ and the second one, but for the asymptotic normality we require a higher moment condition on U_1 . The first part of (A3) is typical for nonparametric smoothing, and is for a bias expansion of the estimator. The second part of (A3) is to deal with those terms involving w_j in the bias expansion. The assumptions (A4)-(A6) are used to obtain the leading bias and variance of the estimator. The last assumption (A7) is also required, along with (A2), for a stochastic expansion of the estimator.

5.2 Proof of Theorem 3.1

Hereafter, $\boldsymbol{\theta}$ denotes the true function. We also let $\boldsymbol{\Theta}$ denote the matrix of the partial derivatives of the true vector function, that is, $\boldsymbol{\Theta}_{jl}(\mathbf{x}, \mathbf{z}) = \partial \theta_j(\mathbf{x}, \mathbf{z}) / \partial x_l$, where θ_j is the j th component function of $\boldsymbol{\theta}$ and x_l is the l th coordinate of $\mathbf{x}$. Define, for a given $(\mathbf{x}, \mathbf{z})$,

$$\tilde{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) = \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z})(\mathbf{u} - \mathbf{x}).$$

The function $\tilde{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v})$ is an approximation of $\boldsymbol{\theta}(\mathbf{u}, \mathbf{v})$ for $\mathbf{u}$ near $\mathbf{x}$ and for $\mathbf{v}$ near $\mathbf{z}$, which is linear in the direction of $\mathbf{x}$, while constant in the direction of $\mathbf{z}$. Define $\mathbf{l}(\mathbf{u}) = (1, \mathbf{u}^\top)^\top$ for $\mathbf{u} \in \mathbb{R}^p$, and

$$\mathbf{G}(\boldsymbol{\alpha}, \mathbf{A}) = f(\mathbf{x}, \mathbf{z}) \int \mathbf{l}(\mathbf{u}) \otimes \boldsymbol{\psi}(\boldsymbol{\alpha} + \mathbf{A}\mathbf{u}|\mathbf{x}, \mathbf{z}) K(\mathbf{u}) d\mathbf{u} \quad (5.1)$$

for $\boldsymbol{\alpha} \in \mathbb{R}^k$ and $\mathbf{A}$ being a $(k \times p)$ -matrix, where $\otimes$ denotes the Kronecker product. Note that $\mathbf{G}$ is a vector of $k(p+1)$ multivariate functions. This is the population version of

$$\begin{aligned} \mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) &:= n^{-1} \sum_{i=1}^n \mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x})) \otimes \mathbf{g}_1 \left(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i) + \boldsymbol{\alpha} + \mathbf{A}\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x}) \right) \\ &\quad \times K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}). \end{aligned}$$

The function $\mathbf{G}_n$ is obtained if we differentiate $\tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) := L_n(\boldsymbol{\theta}(\mathbf{x}, \mathbf{z}) + \boldsymbol{\alpha}, \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z}) + \mathbf{A}\mathbf{H}^{-1})$ with respect to $\boldsymbol{\alpha}$ and $\mathbf{A}$, where L_n is defined at (2.1). The top k entries of $\mathbf{G}_n$ are the partial derivatives $\partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial \alpha_1, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial \alpha_2, \dots, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial \alpha_k$, and the next k entries are $\partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{11}, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{21}, \dots, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{k1}$, and the last k entries are $\partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{1p}, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{2p}, \dots, \partial \tilde{L}_n(\boldsymbol{\alpha}, \mathbf{A}) / \partial A_{kp}$, where we write $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)^\top$ and $\mathbf{A} = (A_{ij})$. Define

$$\hat{\boldsymbol{\alpha}}(\mathbf{x}, \mathbf{z}) = \hat{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z}), \quad \hat{\mathbf{A}}(\mathbf{x}, \mathbf{z}) = \left[\hat{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{z}) - \boldsymbol{\Theta}(\mathbf{x}, \mathbf{z}) \right] \mathbf{H}.$$

Then, $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})$ is the solution of the equation $\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) = \mathbf{0}$.

We claim that, for any compact set $\mathcal{C}$ of $(\boldsymbol{\alpha}, \mathbf{A})$, one has

$$\sup_{(\boldsymbol{\alpha}, \mathbf{A}) \in \mathcal{C}} \|\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) - E\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A})\| = O_p(n^{-1/2} |\mathbf{H}|^{-1/2} (\log n)^{1/2}) \quad (5.2)$$

$$\sup_{(\boldsymbol{\alpha}, \mathbf{A}) \in \mathcal{C}} \|E\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) - \mathbf{G}(\boldsymbol{\alpha}, \mathbf{A})\| = o(1). \quad (5.3)$$

These two properties imply the uniform convergence of $\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A})$ to $\mathbf{G}(\boldsymbol{\alpha}, \mathbf{A})$ in probability over any compact set $\mathcal{C}$. Due to the first part of the assumption (A1), we can conclude that all the entries of $\hat{\boldsymbol{\alpha}}(\mathbf{x}, \mathbf{z})$ and $\hat{\mathbf{A}}(\mathbf{x}, \mathbf{z})$ converge to zero in probability. This enables us to further expand $\mathbf{G}_n(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}}) = \mathbf{0}$ around the solution of $\mathbf{G}(\boldsymbol{\alpha}, \mathbf{A}) = \mathbf{0}$ which are $(\boldsymbol{\alpha}, \mathbf{A}) = (\mathbf{0}, \mathbf{O})$. Define

$$\begin{aligned} \mathbf{J}_n(\boldsymbol{\alpha}, \mathbf{A}) &:= n^{-1} \sum_{i=1}^n [\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x})) \mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x}))^\top] \\ &\quad \otimes \mathbf{g}_2 \left(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i) + \boldsymbol{\alpha} + \mathbf{A}\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x}) \right) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}). \end{aligned}$$

This is obtained by differentiating $\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A})$ with respect to $\boldsymbol{\alpha}$ and $\mathbf{A}$. Let $\mathbf{v}(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})$ denote a $k(p+1)$ -vector obtained by concatenating the entries of $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{A}}$. It is defined by $\mathbf{v}(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})^\top =$

$(\hat{\boldsymbol{\alpha}}^\top, \hat{\mathbf{A}}_1^\top, \dots, \hat{\mathbf{A}}_p^\top)$, where $\hat{\mathbf{A}} = [\hat{\mathbf{A}}_1, \dots, \hat{\mathbf{A}}_p]$. Then, it follows that, for some $(\boldsymbol{\alpha}^*, \mathbf{A}^*)$ such that $\|(\boldsymbol{\alpha}^*, \mathbf{A}^*)\| \leq \|(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})\|$,

$$\mathbf{0} = \mathbf{G}_n(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}}) = \mathbf{G}_n(\mathbf{0}, \mathbf{O}) + \mathbf{J}_n(\boldsymbol{\alpha}^*, \mathbf{A}^*)\mathbf{v}(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}}). \quad (5.4)$$

For $\mathbf{J}_n(\boldsymbol{\alpha}, \mathbf{A})$ we will show that, for any compact set $\mathcal{C}$,

$$\sup_{(\boldsymbol{\alpha}, \mathbf{A}) \in \mathcal{C}} \|\mathbf{J}_n(\boldsymbol{\alpha}, \mathbf{A}) - E\mathbf{J}_n(\boldsymbol{\alpha}, \mathbf{A})\| = o_p(1). \quad (5.5)$$

This entails with the second part of the assumption (A2)

$$\mathbf{J}_n(\boldsymbol{\alpha}^*, \mathbf{A}^*) = E\mathbf{J}_n(\mathbf{0}, \mathbf{O}) + o_p(1). \quad (5.6)$$

To see this, note that the second part of the assumption (A2) implies that for a given $\delta > 0$ there exists $\varepsilon > 0$ such that, for sufficiently large n , $\|E\mathbf{J}_n(\boldsymbol{\alpha}, \mathbf{A}) - E\mathbf{J}_n(\mathbf{0}, \mathbf{O})\| \leq \delta$ for all $(\boldsymbol{\alpha}, \mathbf{A})$ with $\|(\boldsymbol{\alpha}, \mathbf{A})\| \leq \varepsilon$. This and the consistency of $(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})$ together with (5.5) establish (5.6). Define a diagonal matrix $\mathbf{M}$ of dimension $(p+1)$ in such a way that the first entry equals 1 and the rest are all μ_2 . We claim

$$E\mathbf{J}_n(\mathbf{0}, \mathbf{O}) = -[\mathbf{M} \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})]f(\mathbf{x}, \mathbf{z}) + o(1). \quad (5.7)$$

The expansions (5.4), (5.6) and (5.7) give

$$\begin{aligned} \hat{\boldsymbol{\alpha}}(\mathbf{x}, \mathbf{z}) &= [\mathbf{I}_k, \mathbf{O}, \dots, \mathbf{O}] \mathbf{v}(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}}) \\ &= f(\mathbf{x}, \mathbf{z})^{-1} [\mathbf{1}_{p+1}^\top \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})^{-1}] \mathbf{G}_n(\mathbf{0}, \mathbf{O})[1 + o_p(1)], \end{aligned} \quad (5.8)$$

where $\mathbf{1}_{p+1}$ denotes the $(p+1)$ -dimensional unit vector such that $\mathbf{1}_{p+1}^\top = (1, 0, \dots, 0)$.

Now, we derive the first-order properties of $\mathbf{G}_n(\mathbf{0}, \mathbf{O})$. For $\mathbf{Z}^i = \mathbf{z}$, $\Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}) = 1$. For $\mathbf{Z}^i$ with $d(\mathbf{Z}^i, \mathbf{z}) = 1$, we have $\Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}) = w_j$ for some j . Those $\mathbf{Z}^i$ with $d(\mathbf{Z}^i, \mathbf{z}) \geq 2$ have a contribution of order $O_p(w^{*2})$ to $\mathbf{G}_n(\mathbf{0}, \mathbf{O})$, where $w^* = \max_{1 \leq j \leq d} w_j$. Thus,

$$\begin{aligned} \mathbf{G}_n(\mathbf{0}, \mathbf{O}) &= n^{-1} \sum_{i=1}^n \mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x})) \otimes \mathbf{g}_1(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i)) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}) \\ &= n^{-1} \sum_{i=1}^n \mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x})) \otimes \mathbf{g}_1(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i)) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) I(\mathbf{Z}^i = \mathbf{z}) \\ &\quad + n^{-1} \sum_{i=1}^n \mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x})) \otimes \mathbf{g}_1(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i)) K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z}) \\ &\quad \times I[d(\mathbf{Z}^i, \mathbf{z}) = 1] + O_p(w^{*2}) \\ &\stackrel{\text{let}}{=} \mathbf{T}_1 + \mathbf{T}_2 + O_p(w^{*2}). \end{aligned} \quad (5.9)$$

The expected value of the first term in (5.9) has the following expansion due to the second part of the assumption (A1) and the assumptions (A2)–(A5):

$$\begin{aligned}
E(\mathbf{T}_1) &= E [\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x})) \otimes \mathbf{g}_2(Y, \boldsymbol{\theta}(\mathbf{X}, \mathbf{z}))] [\tilde{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{z}) - \boldsymbol{\theta}(\mathbf{X}, \mathbf{z})] K_{\mathbf{H}}(\mathbf{X} - \mathbf{x}) I(\mathbf{Z} = \mathbf{z}) \\
&\quad + o(\text{tr}(\mathbf{H}^2)) \\
&= E [\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x})) \otimes \boldsymbol{\rho}(\mathbf{X}, \mathbf{z})] [\boldsymbol{\theta}(\mathbf{X}, \mathbf{z}) - \tilde{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{z})] K_{\mathbf{H}}(\mathbf{X} - \mathbf{x}) I(\mathbf{Z} = \mathbf{z}) \\
&\quad + o(\text{tr}(\mathbf{H}^2)) \\
&= \frac{1}{2} f(\mathbf{x}, \mathbf{z}) \int [\mathbf{l}(\mathbf{u}) \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})] [\mathbf{u}^\top \mathbf{H} \boldsymbol{\theta}''_j(\mathbf{x}, \mathbf{z}) \mathbf{H} \mathbf{u}] K(\mathbf{u}) d\mathbf{u} + o(\text{tr}(\mathbf{H}^2)).
\end{aligned}$$

By the properties of the multivariate kernel K , we can further approximate $E(\mathbf{T}_1)$ by

$$E(\mathbf{T}_1) = \frac{1}{2} \mu_2 f(\mathbf{x}, \mathbf{z}) [\mathbf{1}_{p+1} \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z})] \boldsymbol{\beta}_{\mathbf{H}}(\mathbf{x}, \mathbf{z}) + o(\text{tr}(\mathbf{H}^2)),$$

where $\boldsymbol{\beta}_{\mathbf{H}}(\mathbf{x}, \mathbf{z})$ is k -vector whose j th entry equals $\text{tr}(\boldsymbol{\theta}''_j(\mathbf{x}, \mathbf{z}) \mathbf{H}^2)$. One can similarly get an approximation of $\text{var}(\mathbf{T}_1)$. In fact,

$$\text{var}(\mathbf{T}_1) = n^{-1} |\mathbf{H}|^{-1} f(\mathbf{x}, \mathbf{z}) \cdot \mathbf{D} \otimes \boldsymbol{\tau}(\mathbf{x}, \mathbf{z}) + o(n^{-1} |\mathbf{H}|^{-1}),$$

where $\mathbf{D}$ is a $(p+1)$ -dimensional diagonal matrix whose first diagonal entry equals $\int K^2(\mathbf{u}) d\mathbf{u}$ and the next p diagonal entries are $\int u_j^2 K^2(\mathbf{u}) d\mathbf{u}$, $1 \leq j \leq p$.

Next, we look into the term $\mathbf{T}_2$. This term contributes only $E(\mathbf{T}_2)$ to the first-order properties of $\mathbf{G}_n(\mathbf{0}, \mathbf{O})$ since $\text{var}(\mathbf{T}_2)$ is negligible in comparison with $\text{var}(\mathbf{T}_1)$ because of the additional factors w_j that go to zero as n tends to infinity. For an expansion of $E(\mathbf{T}_2)$, we note that the following approximation holds due to the second part of (A3): uniformly for $\mathbf{v} \in \mathcal{D}$,

$$\boldsymbol{\theta}(\mathbf{u}, \mathbf{v}) - \tilde{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) = [\boldsymbol{\theta}(\mathbf{x}, \mathbf{v}) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z})] + O(\|\mathbf{u} - \mathbf{x}\|)$$

for $\mathbf{u}$ near $\mathbf{x}$. With this and using the assumptions (A4) and (A5) we get

$$\begin{aligned}
E(\mathbf{T}_2) &= E [\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x})) \otimes \boldsymbol{\rho}(\mathbf{X}, \mathbf{Z})] [\boldsymbol{\theta}(\mathbf{X}, \mathbf{Z}) - \tilde{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z})] K_{\mathbf{H}}(\mathbf{X} - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}, \mathbf{z}) \\
&\quad \times I[d(\mathbf{Z}, \mathbf{z}) = 1] + o(w^*) \\
&= E [\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x})) \otimes \boldsymbol{\rho}(\mathbf{X}, \mathbf{Z})] [\boldsymbol{\theta}(\mathbf{x}, \mathbf{Z}) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z})] K_{\mathbf{H}}(\mathbf{X} - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}, \mathbf{z}) \\
&\quad \times I[d(\mathbf{Z}, \mathbf{z}) = 1] + o(w^*) \\
&= \sum_{\mathbf{z}': d(\mathbf{z}', \mathbf{z})=1} f(\mathbf{x}, \mathbf{z}') [\mathbf{1}_{p+1} \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z}')] [\boldsymbol{\theta}(\mathbf{x}, \mathbf{z}') - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z})] \Lambda_{\mathbf{w}}(\mathbf{z}', \mathbf{z}) + o(w^*) \\
&= \sum_{j=1}^d w_j \sum_{\substack{\mathbf{z}'_j \neq \mathbf{z}_j, \mathbf{z}'_j \in \mathcal{D}_j}} f(\mathbf{x}, \mathbf{z}_{-j}, \mathbf{z}'_j) [\mathbf{1}_{p+1} \otimes \boldsymbol{\rho}(\mathbf{x}, \mathbf{z}_{-j}, \mathbf{z}'_j)] [\boldsymbol{\theta}(\mathbf{x}, \mathbf{z}_{-j}, \mathbf{z}'_j) - \boldsymbol{\theta}(\mathbf{x}, \mathbf{z})] \\
&\quad + o(w^*).
\end{aligned}$$

Asymptotic normality of $\mathbf{T}_1$ follows from a standard technique and the first part of the assumption (A2). The theorem now follows from some basic properties of kronecker products. It remains to prove (5.2), (5.3), (5.5) and (5.7). Among them, (5.7) can be proved similarly as in the derivation of the expansion for $E(\mathbf{T}_1)$.

We prove (5.2) first. We write simply $\mathbf{l}^i$ for $\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x}))$, $\mathbf{g}_1^i(\boldsymbol{\alpha}, \mathbf{A})$ for $\mathbf{g}_1(Y^i, \tilde{\boldsymbol{\theta}}(\mathbf{X}^i, \mathbf{Z}^i) + \boldsymbol{\alpha} + \mathbf{A}\mathbf{H}^{-1}(\mathbf{X}^i - \mathbf{x}))$, K^i for $K_{\mathbf{H}}(\mathbf{X}^i - \mathbf{x})$ and Λ^i for $\Lambda_{\mathbf{w}}(\mathbf{Z}^i, \mathbf{z})$. Define $\boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A}) = (\mathbf{l}^i \otimes \mathbf{g}_1^i(\boldsymbol{\alpha}, \mathbf{A}))K^i\Lambda^i$. Then, we can write $\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) = n^{-1} \sum_{i=1}^n \boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A})$. We want to get an exponential bound for a large deviation of the centered $\sqrt{n|\mathbf{H}|/\log n}\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A})$ for each fixed $(\boldsymbol{\alpha}, \mathbf{A})$. Since $\boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A})$ are not bounded, we employ a truncation technique. Since $\Lambda^i \leq 1$ for all $1 \leq i \leq n$ and from the first part of the assumption (A2), we obtain that for any compact set $\mathcal{C}$

$$\begin{aligned} & \sup_{(\boldsymbol{\alpha}, \mathbf{A}) \in \mathcal{C}} \left\| n^{-1} \sum_{i=1}^n \boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A}) I(\|\boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A})\| > \sqrt{n}) \right\| \\ & \leq Cn^{-1} \sum_{i=1}^n U_1(Y^i) I[U_1(Y^i) > C\sqrt{n}] K^i. \end{aligned} \quad (5.10)$$

The right hand side of (5.10) has the expectation of the magnitude $O(n^{-1/2})$ due to the fact that the conditional second moment of $U_1(Y)$ given $\mathbf{X} = \mathbf{u}$ is bounded locally uniformly for $\mathbf{u}$ around $\mathbf{x}$, see the first part of (A2). This implies that the left hand side of (5.10) is of order $O_p(n^{-1/2})$. Similarly, we also get $E[\boldsymbol{\xi}(\boldsymbol{\alpha}, \mathbf{A}) I(\|\boldsymbol{\xi}(\boldsymbol{\alpha}, \mathbf{A})\| > \sqrt{n})] = O(n^{-1/2})$ uniformly for $(\boldsymbol{\alpha}, \mathbf{A})$ in any compact set. These considerations reduce the proof of (5.2) to that for the truncated version $\mathbf{G}_n^*(\boldsymbol{\alpha}, \mathbf{A}) := n^{-1} \sum_{i=1}^n \boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A}) I(\|\boldsymbol{\xi}^i(\boldsymbol{\alpha}, \mathbf{A})\| \leq \sqrt{n})$. By a simple application of Markov inequality and since $|\mathbf{H}|E\|\boldsymbol{\xi}(\boldsymbol{\alpha}, \mathbf{A})\|^2$ is bounded, say by c , from the first part of the assumption (A2), we get

$$P\left[\|\mathbf{G}_n^*(\boldsymbol{\alpha}, \mathbf{A}) - E\mathbf{G}_n^*(\boldsymbol{\alpha}, \mathbf{A})\| > M\sqrt{(\log n)/(n|\mathbf{H}|)}\right] \leq n^{c-M} \quad (5.11)$$

for any fixed $\boldsymbol{\alpha}$ and $\mathbf{A}$. Since $\mathbf{G}_n$ is Lipschitz continuous of order 1 with a Lipschitz constant $O_p(1)$ by the second part of the assumption (A2), the exponential bound (5.11) concludes the proof of (5.2).

Next, we prove (5.3). By the assumption (A7), we obtain

$$\begin{aligned} E\mathbf{G}_n(\boldsymbol{\alpha}, \mathbf{A}) &= E\left[\mathbf{l}(\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x})) \otimes \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z}) - \boldsymbol{\theta}(\mathbf{X}, \mathbf{Z}) + \boldsymbol{\alpha} + \mathbf{A}\mathbf{H}^{-1}(\mathbf{X} - \mathbf{x}) | \mathbf{X}, \mathbf{Z})\right] \\ &\quad \times K_{\mathbf{H}}(\mathbf{X} - \mathbf{x}) \Lambda_{\mathbf{w}}(\mathbf{Z}, \mathbf{z}) \\ &= \int \left[\mathbf{l}(\mathbf{u}) \otimes \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}(\mathbf{x} + \mathbf{H}\mathbf{u}, \mathbf{z}) - \boldsymbol{\theta}(\mathbf{x} + \mathbf{H}\mathbf{u}, \mathbf{z}) + \boldsymbol{\alpha} + \mathbf{A}\mathbf{u} | \mathbf{x} + \mathbf{H}\mathbf{u}, \mathbf{z})\right] \\ &\quad \times f(\mathbf{x} + \mathbf{H}\mathbf{u}, \mathbf{z}) K(\mathbf{u}) d\mathbf{u} + O(w^*) \\ &= \mathbf{G}(\boldsymbol{\alpha}, \mathbf{A}) + o(1) \end{aligned}$$

uniformly for $(\boldsymbol{\alpha}, \mathbf{A})$ in any compact set. This completes the proof of (5.3). The proof of (5.5) is similar to that of (5.2). For this proof, one may use continuity of $\mathbf{g}_2(y, \boldsymbol{\theta})$ in $\boldsymbol{\theta}$ and the following exponential inequality for the truncated version of $\mathbf{J}_n$, denoted by $\mathbf{J}_n^*$, constructed in the same way as $\mathbf{G}_n^*$: for any $\varepsilon > 0$ it holds that

$$P\left[\|\mathbf{J}_n^*(\boldsymbol{\alpha}, \mathbf{A}) - E\mathbf{J}_n^*(\boldsymbol{\alpha}, \mathbf{A})\| > \varepsilon\right] \leq n^c e^{-\varepsilon \sqrt{n|\mathbf{H}|\log n}}$$

for any fixed $\boldsymbol{\alpha}$ and $\mathbf{A}$, where c is the same positive constant as at (5.11).

6 Conclusions

In this paper the theory for local maximum likelihood estimation in the presence of categorical data has been established. These techniques are particularly useful for models where the information contained in the likelihood is useful for identifying and estimating the model. This is the case when we are given the conditional density of the variable of interest Y given $\mathbf{X} = \mathbf{x}$ that can be written as $\text{pdf}(y|\mathbf{x}) = g(y, \boldsymbol{\theta}(\mathbf{x}))$, where $g(\cdot)$ is a known function and $\boldsymbol{\theta}(\mathbf{x}) \in \mathbb{R}^k$ is a set of k unknown functions of $\mathbf{x}$. The aim of local maximum likelihood methods is to use local polynomial approximations of these functions and then to maximize locally the likelihood function.

Examples of such econometric models include probit and logit models, truncated regression models, stochastic frontier models, etc. The paper develops the methodology based on using discrete kernels proposed by Racine and Li (2004), and inspired by Aitchison and Aitken (1976). In a sense, our theory is parallel to the theory provided by Racine and Li (2004), who justified the use of their kernels to handle categorical variables among regressors in non-parametric regression estimation based on the local least squares estimator. Indeed, we note that many of the asymptotic results found in Racine and Li (2004) for the local least squares estimator with their kernel are pertinent to the local likelihood estimator.

In particular, under mild regularity conditions, the local likelihood estimator with Li-Racine kernels preserves consistency and asymptotic normality properties. Importantly, its first order properties depend on the weights of the discrete kernel, but only through the bias. Moreover, if we assume that the discrete bandwidth is of order $o(\text{tr}(\mathbf{H}^2))$ (where $\mathbf{H}$ is the matrix of bandwidths for continuous regressors), then the leading term of the bias that involves the discrete kernel becomes also negligible. Our theorem also shows dependency of the bias on the curvature of the unknown functions. Remarkably, and very analogous to the case of local least-squares estimators, the rate of convergence does not depend on the presence of the discrete variables. In other words, there is no ‘‘curse of dimensionality’’ for the local likelihood estimator with respect to the discrete regressors. This is a very important

property, giving to local likelihood estimator with Li-Racine kernels a clear advantage over separate application of local likelihood for each value of the discrete variable. On the other hand, our theorem coincides with the classical result on the local likelihood when there are no discrete variables but only continuous ones.

It is also worth noting that, as was shown in Park et al. (2008) for the truncated regression model, using lower order polynomial for the variance function deteriorates the rate of convergence of the regression function estimator and that, in the case of constant variance function (a model in the spirit of Severini and Wong, 1992), one can use a 2-stage procedure that would give a $\sqrt{n}$ -consistent estimator of the variance (see Park et al. (2008) for details). These two results remain valid in the current context we considered here.

We provided theory under fairly mild conditions on the data generating process, on the kernels and on the bandwidth. Yet, of course, as with other smoothing estimators, the choice of the bandwidths in practice would be critical for our estimator too. For making this choice in practice, we suggest joint selection of both continuous and discrete bandwidths by maximizing a leave-one-out cross validation (CV) criterion that is based on the estimated log-likelihood. It is also worth noting that, as in Racine and Li (2004), it remains true that, if there is no continuous variable, the CV choice of $\mathbf{w}$ will converge to zero at the rate n^{-1} . Moreover, if the “working parametric model” is true, then one may get the parametric rate of convergence by letting $h \rightarrow \infty$.

We also illustrated our method in various simulated scenarios and the results indicate a great flexibility of the approach and good performances in complex scenarios, even with moderate sample sizes (e.g., $n = 100$).

Natural extension of this work goes towards the statistical testing of various hypotheses, including testing about the sign or size of the impact (marginal effects) from continuous and discrete regressors included into the model, or relevance of any of these regressors. Another extension is the accommodation of time series and panel data contexts.

References

- [1] Aigner, D. J., C. A. K Lovell and P. Schmidt (1977), Formulation and estimation of stochastic frontier models, *Journal of Econometrics*, 6, 21-37.
- [2] Aitchison, J. and C. G. G. Aitken (1976), Multivariate binary discrimination by the kernel method, *Biometrika*, 63(3), 413–420.

- [3] Eguchi, S., T. Y. Kim and B. U. Park (2003), Local Likelihood method: a bridge over parametric and nonparametric regression, *Journal of Nonparametric Statistics*, 15(6), 665–683.
- [4] Fan, J., N. E. Heckman and M. P. Wand (1995), Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions, *Journal of the American Statistical Association*, 90(429), 141–150.
- [5] Frölich, M. (2006), Non-parametric regression for binary dependent variables, *Econometrics Journal*, 9, 511–540.
- [6] Kumbhakar, S. C., B. U. Park, L. Simar and E. G. Tsionas (2007), Nonparametric stochastic frontiers: a local likelihood approach, *Journal of Econometrics*, 137(1), 1–27.
- [7] Li, Q. and J. S. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.
- [8] Park, B. U., L. Simar and V. Zelenyuk (2008), Local Likelihood Estimation of Truncated Regression and its Partial Derivatives: Theory and Application, *Journal of Econometrics*, 146(1), 185–198.
- [9] 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.
- [10] Severini, T. A. and W. H. Wong (1992), Profile Likelihood and Conditionally Parametric Models, *Annals of Statistics*, 20(4), 1768–1802.