Local Composite Quantile Regression Smoothing: Flexible Data Structure and Cross-validation

Xiao Huang *
Kennesaw State University

Zhongjian Lin†
Emory University

March 5, 2018

Abstract

In this paper, we study the local composite quantile regression estimator for mixed categorical and continuous data. The local composite quantile estimator is an efficient and safe alternative to the local polynomial method and has been well-studied for continuous covariates. The generalization of the local composite quantile regression estimator to a flexible data structure is appealing to practitioners as empirical studies often encounter categorical data. Furthermore, we study the theoretical properties of the cross-validated bandwidth selection for the local composite quantile estimator. Under mild conditions, we derive the rate of convergence of the cross-validated smoothing parameters to their optimal benchmark values for both categorical and continuous covariates. Monte Carlo experiments show that the proposed estimator may have a large efficiency gain compared to the local linear estimator.

Keywords: Local Composite Quantile Regression, Mixed Categorical and Continuous Data, Cross-validation, Efficiency.
JEL Classifications: C14; C21.

1 Introduction

Nonparametric methodology is a flexible tool to study the relationship among variables and avoids some potential misspecification problems. There have been extensive studies on various nonparametric estimation methods when the covariates are continuous, see Härdle (1990); Härdle and Linton (1994); Pagan and Ullah (1999); Li and Racine (2007); Chen (2007); Ichimura and Todd (2007); Horowitz (2009); Henderson and Parmeter (2015). However, economic analyses often encounter categorical data, such as age, gender, race, etc., which prevents the boom of empirical applications using nonparametric techniques. Racine and Li (2004); Li and Racine (2004) generalize the local linear smoothing to allow for both categorical and continuous covariates.

*Corresponding author. Email: xhuang3@kennesaw.edu.
†Email: zhongjian.lin@emory.edu.

The local composite quantile regression (CQR) smoothing technique is an efficient and safe alternative to the local polynomial method and has been well-studied for continuous covariates. Kai, Li, and Zou (2010) derive the asymptotic bias, variance, and normality of the local CQR estimator for continuous covariates and document the relative efficiency compared to the local polynomial method for non-normal errors. The local composite quantile regression smoothing technique is also extended for semiparametric models, e.g. Kai, Li, and Zou (2011) for varyling coefficient partially linear model, Jiang, Zhou, Qian, and Shao (2012); Fan, Härdle, Wang, and Zhu (2013) for single-index model. Li and Li (2016) generalize the local composite quantile regression smoothing to Harris recurrent Markov processes, which include both stationary and nonstationary data. Zhao and Xiao (2014) provide a weighted version of the local composite quantile regression smoothing and show that its variance achieves the Cramér-Rao lower bound. Kong and Xia (2014) combine the composite quantile smoothing technique with a dimension reduction method for high-dimensional analysis.

In this paper, we aim to bridge the gap of the local CQR smoother for a flexible data structure. Consider the nonparametric regression model defined by

$$Y_i = g(X_i) + \sigma(X_i)\varepsilon_i, i = 1, \cdots, n, \tag{1}$$

where $X_i \equiv (X_i^c, X_i^d)$ contains both continuous and categorical variables, $g(\cdot)$ is the mean function, and $\sigma^2(\cdot)$ is the conditional variance. The mixed categorical and continuous data structure has been extensively studied in the literature, see, for example, Li and Racine (2003); Racine and Li (2004); Li and Racine (2004); Sun (2006); Hsiao, Li, and Racine (2007); Li and Racine (2008); Li, Racine, and Wooldridge (2008); Li, Maasoumi, and Racine (2009); Li, Racine, and Wooldridge (2009); Ouyang, Li, and Racine (2009); Li and Racine (2010) to name only a few. In this paper, we derive the asymptotic bias, variance, and normality of the local composite quantile smoother for mixed categorical and continuous covariates.

It is well-known that nonparametric kernel estimators are sensitive to the choice of the smoothing parameter. Undersmoothing or oversmoothing leads to an unsatisfactory large variance or large bias. Therefore, a data-driven optimal bandwidth selection is necessary in applied nonparametric settings. We propose least squares cross-validation to select optimal bandwidths for both categorical and continuous covariates in the local composite quantile regression smoother. We derive the rate of convergence of the cross-validated bandwidths to their optimal benchmark values. In our Monte Carlo experiments, we find that the cross-validated bandwidths for categorical covariates in the local CQR smoother are around three times larger than those selected by cross-validation in the local linear method [Li and Racine (2004)].

This paper unfolds as follows. Section 2 presents the local CQR estimator for the nonparametric model with mixed categorical and continuous data. The asymptotic bias, variance, and normality of the local CQR estimator are also derived. Section 3 presents the cross-validated bandwidth selection for the local CQR estimator with mixed data and the convergence rate of the selected bandwidths to their theoretical benchmark values. We illustrate the relative efficiency of the cross-validated local CQR estimator to the corresponding local linear estimation with three Monte Carlo experiments in Section 4. Section 5 concludes.

2 Local CQR Smoothing for Nonparametric Regression with Mixed Data

We consider a nonparametric regression model in Equation (1) with both categorical and continuous regressors, where ε_i is the error term with mean 0 and variance 1, and $X_i \equiv (X_i^c, X_i^d)$ with X_i^d denoting an $r \times 1$ vector of categorical covariates and X_i^c denoting a $p \times 1$ vector of continuous regressors. We use $X_{t,i}^d$ to denote the tth component of X_i^d and we assume that $X_{t,i}^d \in \{0, 1, \cdots, C-1\}$ for $t = 1, \cdots, r$ and $C \geq 2$. $g(\cdot)$ has an unknown functional form. We use $f(x) = f(x^c, x^d)$ to denote the joint density function of X_i. Define the partial derivative of $g(X)$ with respect to X^c:

$$\beta(X) \equiv \frac{\partial g(X)}{\partial X^c} \equiv g'(X), \tag{2}$$

where we use prime to denote the partial derivative of the function with respect to a continuous regressor from here on.

Define $d_{ix} = \sum_{t=1}^{r} 1\{X_{t,i}^d \neq x_t^d\}$, which equals the number of disagreeing components between X_i^d and x^d. We use the following kernel function for categorical variables with bandwidth λ:

$$L_\lambda(X_i^d, x^d) = \lambda^{d_{ix}}.$$

Denote $w(\cdot)$ as the kernel function associated with the continuous variables X^c and h as the corresponding bandwidth. Denote $W(x) = \Pi_{k=1}^{p} w(x)$. The kernel function for the mixed regressors becomes

$$K_{ix} \equiv W_{ix} L_{\lambda,ix} = W\Big(\frac{X_i^c - x^c}{h}\Big) L_\lambda\Big(X_i^d, x^d\Big).$$

We propose the local linear CQR smoother as an efficient alternative to the local linear regression estimator. Denote the CQR loss function as

$$\sum_{k=1}^{q} \sum_{i=1}^{n} \rho_{\tau_k}\Big(Y_i - a_k - b(X_i^c - x^c)\Big) K_{ix}, \tag{3}$$

where $\rho_{\tau_k}(u) \equiv u\Big(\tau_k - 1\{u < 0\}\Big)$ is the check function in quantile regression. Denote the minimizer of Equation (3) with respect to $(a_1, \cdots, a_q, b)$ as $(\hat{a}_1, \cdots, \hat{a}_q, \hat{b})$. We then have the local linear CQR estimators of $g(x)$ and $\beta(x)$ as

$$\hat{g}(x) = \frac{1}{q}\sum_{k=1}^{q}\hat{a}_k,$$
$$\hat{\beta}(x) = \hat{b}.$$

Note that Equation (3) treats the continuous regressor X^c in a local linear fashion and the categorical regressor X^d in a local constant fashion. The number of quantiles, q, is essentially another tuning parameter and is chosen by the researcher. We show in simulation that there can be considerable gain by choosing a relatively small q. In this section, we investigate the asymptotic properties of the proposed local linear CQR smoother for a flexible data structure. To derive the asymptotic results, we make the following assumptions.

Assumption 1. (i) The support of X, $\mathcal{X} \equiv \mathcal{X}^c \times \mathcal{X}^d$, is compact; (ii) $g(X)$ is twice differentiable with respect to X^c in $\mathcal{X}^c$; (iii) $(X_i, Y_i), i = 1, 2, \cdots, n$ are i.i.d. and $\mathbb{E}(Y_i|X_i) = g(X_i)$ almost surely; (iv)Letting $g''(X)$ denote the Hessian of g with respect to X^c, then $\int g''(X)^2 f(X)dX > 0$.

Assumption 2. (i) $f(X)$ is twice differentiable with respect to X^c and $E\left[f^4(X)\right] < \infty$; (ii) The density of X, $f(X)$, is positive.

Denote

$$\mu_j = \int u^j w(u)du \qquad \text{and} \qquad \nu_j = \int u^j w^2(u)du, \qquad j = 0, 1, \cdots. \tag{4}$$

Assumption 3. The kernel function $w(X)$ has compact support and is bounded and symmetric such that $\int vv^T W(v)dv = \mu_2 I_p < \infty$ and $\int vv^T W^2(v)dv = v_2 I_p < \infty$, where I_p is a $p \times p$ identity matrix.

Assumption 4. (i) $\sigma^2(\cdot)$ is twice differentiable with respect to X^c; (ii) $\mathbb{E}[\sigma^4(X)] < \infty$.

Assumption 5. ε_i has a symmetric distribution with continuous and positive density.

Remark 1. Assumption 1 (i), (ii) and Assumption 2 are standard conditions used in the nonparametric literature. Assumption 1 (iii) is essentially an exogenous condition. Assumption 1 (iv) indicates that all covariates are relevant. In nonparametric literature, there is a technique of smoothing out irrelevant variables by data-driven cross-validation bandwidth selections, i.e. $\hat{h} \to \infty$ and $\hat{\lambda} = 1$. See Hall, Racine, and Li (2004); Hall, Li, and Racine (2007) for details. In writing Equation (4), we also implicitly assume that μ_j and v_j exist and can be viewed as moments of individual kernels $w(u)$ and $w^2(u)$. Assumption 3 further assumes that kernel moments for different elements of u are zero. This assumption simplifies the presentation of the results and can is commonly found in the nonparametric literature, e.g., Ruppert and Wand (1994). We assume $\sigma^2(X)$ is twice differentiable in X^c as second-order Taylor series expansion is needed in obtaining leading terms of σ at the point of interest. The symmetry requirement in Assumption 5 is used in the proof of Theorem 2 to obtain the explicit bias expression and it ensures the leading term in the bias is $O(h^2 + \lambda)$ which vanishes as $n \to \infty$.

Let $F_\varepsilon(\cdot)$ and $f_\varepsilon(\cdot)$ be the CDF and PDF of the error term ε, respectively. Define

$$R_1(q) = \frac{1}{q^2} \sum_{k=1}^{q} \sum_{k'=1}^{q} \frac{\tau_{kk'}}{f_\varepsilon(c_k) f_\varepsilon(c_{k'})}, \tag{5}$$

where $c_k = F_\varepsilon^{-1}(\tau_k)$ and $\tau_{kk'} = \tau_k \wedge \tau_{k'} - \tau_k \tau_{k'}$.

Define, for $k = 1, \cdots, q$,

$$\begin{aligned}
u_k &\equiv \sqrt{nh^p}\{a_k - g(x) - \sigma(x) c_k\}, \\
v &\equiv \sqrt{nh^{p+2}}[b - \beta(x)], \\
\Delta_{i,k} &\equiv \frac{u_k}{\sqrt{nh^p}} + \frac{v^T (X_i^c - x^c)}{\sqrt{nh^{p+2}}}, \\
r_i &\equiv g(X_i) - g(x) - \beta(x)(X_i^c - x^c), \\
d_{i,k} &\equiv c_k[\sigma(X_i) - \sigma(x)] + r_i, \\
\eta_{i,k}^* &\equiv 1\{\varepsilon_i \le c_k - d_{i,k}/\sigma(X_i)\} - \tau_k.
\end{aligned}$$

Denote $\theta = (u_1, \cdots, u_q, v^T)^T$.

Proposition 1. *Minimizing Equation (3) is equivalent to minimizing*

$$L_n(\theta) = \sum_{k=1}^{q} u_k \left(\sum_{i=1}^{n} \frac{K_{ix} \eta_{i,k}^*}{\sqrt{nh^p}} \right) + v^T \left(\sum_{k=1}^{q} \sum_{i=1}^{n} \frac{K_{ix}(X_i^c - x^c) \eta_{i,k}^*}{\sqrt{nh^{p+2}}} \right) + \sum_{k=1}^{q} B_{n,k}(\theta),$$

with respect to θ, where

$$B_{n,k}(\theta) = \sum_{i=1}^{n} K_{ix} \int_0^{\Delta_{i,k}} \Big[1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le w\} - 1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le 0\} \Big] dw.$$

Proof. See Appendix B. □

Let $S_{n,11}$ be a $q \times q$ diagonal matrix with diagonal elements $\frac{f_\varepsilon(c_k)}{nh^p} \sum_{i=1}^n \frac{K_{ix}}{\sigma(X_i)}, k = 1, \cdots, q$, $S_{n,12}$ be a $q \times p$ matrix with k-th row $\frac{f_\varepsilon(c_k)}{nh^{p+1}} \sum_{i=1}^n \frac{K_{ix}(X_i^c - x^c)^T}{\sigma(X_i)}, k = 1, \cdots, q$, and $S_{n,22}$ be a $p \times p$ matrix $\frac{1}{nh^{p+2}} \sum_{k=1}^q f_\varepsilon(c_k) \sum_{i=1}^n \frac{K_{ix}(X_i^c - x^c)(X_i^c - x^c)^T}{\sigma(X_i)}$.

Denote

$$S_n = \begin{pmatrix} S_{n,11} & S_{n,12} \\ S_{n,12}^T & S_{n,22} \end{pmatrix}.$$

Let $W_n^* = (w_{11}^*, \cdots, w_{1q}^*; w_{21})^T$ with

$$w_{1k}^* = \frac{1}{\sqrt{nh^p}} \sum_{i=1}^{n} K_{ix} \eta_{i,k}^*, k = 1, \cdots, q$$

and

$$w_{21}^* = \frac{1}{\sqrt{nh^{p+2}}} \sum_{k=1}^{q} \sum_{i=1}^{n} K_{ix} (X_i^c - x^c) \eta_{i,k}^*.$$

Proposition 2. *Under Assumptions 1 to 5, $L_n(\theta) = \frac{1}{2}\theta^T S_n \theta + (W_n^*)^T \theta + o_p(1)$.*

Proof. See Appendix B. □

Let

$$S = \begin{pmatrix} f(c_1) & 0 & \cdots & 0 & 0 \\ 0 & f(c_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & f(c_q) & 0 \\ 0 & 0 & \cdots & 0 & \mu_2 \sum_{k=1}^{q} f(c_k) \end{pmatrix}.$$

Similarly, let Σ_{11} be a $q \times q$ matrix with (k, k')-element $\nu_0 \tau_{kk'}$, and Σ_{22} be $(\sum_{k,k'=1}^{q} \tau_{kk'})\nu_2$. Define

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

Denote $\eta_{i,k} = 1\{\varepsilon_i \le c_k\} - \tau_k$ and $W_n = (w_{11}, \cdots, w_{1q}, w_{21}^T)^T$ with

$$w_{1k} = \frac{1}{\sqrt{nh^p}} \sum_{i=1}^{n} K_{ix} \eta_{i,k},$$

and

$$w_{21} = \frac{1}{\sqrt{nh^{p+2}}} \sum_{k=1}^{q} \sum_{i=1}^{n} K_{ix} (X_i^x - x^c) \eta_{i,k}.$$

Theorem 1. *Under Assumptions 1 to 5, we have*

$$\hat{\theta}_n + \frac{\sigma(x)}{f(x)} S^{-1} \mathbb{E}(W_n^* | X) \xrightarrow{d} MVN\Big(0, \frac{\sigma^2(x)}{f(x)} S^{-1} \Sigma S^{-1}\Big). \tag{6}$$

Proof. See Appendix B. □

Theorem 1 gives the asymptotic property of $\hat{\theta}_n$, which is used to derive the asymptotic property of $\hat{g}(X)$ in Theorem 2. The bias and variance in Theorem 1 can be used to directly construct the bias and variance in Theorem 2.

Theorem 2. *Under Assumptions 1-5, the asymptotic conditional bias and variance of the local* *linear CQR estimator $\hat{g}(x)$ are given by*

$$Bias(\hat{g}(x)|X) = \left[h^2\frac{tr[\beta'(x)]\mu_2}{2} + \lambda \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \frac{\sigma(x)f(x^c,\tilde{x}^d)[g(x^c,\tilde{x}^d) - g(x)]}{f(x)\sigma(x^c,\tilde{x}^d)}\right](1 + o_p(1)),$$

$$Var(\hat{g}(x)|X) = \frac{1}{nh^p}\frac{\sigma^2(x)}{f(x)}\nu_0 R_1(q)(1 + o_p(1)).$$

Furthermore, conditioning on X, we have

$$\sqrt{nh^p}\left[\hat{g}(x) - g(x) - h^2\frac{tr[\beta'(x)]\mu_2}{2} - \lambda \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \frac{\sigma(x)f(x^c,\tilde{x}^d)[g(x^c,\tilde{x}^d) - g(x)]}{f(x)\sigma(x^c,\tilde{x}^d)}\right] \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{f(x)}\nu_0 R_1(q)\right). \tag{7}$$

Proof. See Appendix B. □

Remark 2. From Theorem 2, the mean squared error (MSE), omitting the smaller order terms, is given by

$$\text{MSE}(x) = (C_1 h^2 + C_2\lambda)^2 + \frac{C_3}{nh^p}, \tag{8}$$

where C_1, C_2, and C_3 are corresponding coefficients in the bias and variance expressions. Minimizing Equation (8) w.r.t. both h and λ will not give the optimal bandwidth h as we can choose λ such that the first term on the r.h.s. of Equation (8) is zero with no solution for h. Adding higher order terms such as $O(\lambda^2)$ and $O(\lambda h)$ in the bias formula may partially solve this problem by expressing the optimal λ as a function of h and solve a higher order polynomial in h in MSE. However, a closed-form solution for optimal h is difficult to obtain. Instead, minimizing the leading terms of the cross-validation objective function provides closed-form solutions to h and λ, which makes the cross-validation in Section 3 not only numerically appealing but also analytically tractable.

Remark 3. The first term in the bias expression is a standard result from the local linear regression. This part of bias essentially measures the impact of curvature in $g(x)$ due to the continuous variables in x. Hence, more curvature will lead to larger diagonal elements in $\beta'(x)$, a larger trace of the Hessian matrix of $g(x)$, and a larger bias. The second term is the additional bias due to the presence of categorical regressors. Categorical variables cause jumps in $g(x)$. The term $g(x^c, \tilde{x}^d) - g(x)$ and the summation over $\tilde{x}^d$ suggests that more jumps necessarily lead to a larger bias. This jump-related bias is also documented in Theorem 2.1 in Racine and Li (2004). Finally, we notice that there is no derivative of $f(x)$ in the bias expression, which implies that despite the presence of categorical variables, the CQR estimator preserves the design-adaptive property of the local least estimator.

3 Cross-validation Bandwidth Selection

In this section, we study the selection of bandwidth based on least squares cross-validation. The general procedure for bandwidth selection with mixed regressors is discussed in Racine and Li (2004) and Li and Racine (2004). We extend the work in Racine and Li (2004) and Li and Racine (2004) to the case of the local composite quantile regression. The form of the local linear CQR estimator differs from the local constant/linear estimator in Racine and Li (2004); Li and Racine (2004), leading to some differences in the proof although the convergence results for bandwidths are similar.

From Equation (6), it follows that

$$\begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_q \end{pmatrix} = -\frac{\sigma(x)}{f(x)} \left(S^{-1}\right)_{11} W_{1n}^*, \tag{9}$$

where $\left(S^{-1}\right)_{11}$ is a diagonal matrix, the upper-left block matrix of S^{-1}, and W_{1n}^* is the first q elements of W_n^*. In Equation (9), we also omit the $o_p(1)$ term in Equation (6) as it will not affect the analysis of cross-validation.

Substitute Equation (9) into $\hat{g}(x)$ to obtain

$$\begin{aligned} \hat{g}(x) &= -\frac{1}{q\sqrt{nh^p}} \frac{\sigma(x)}{f(x)} \iota_q^T \left(S^{-1}\right)_{11} W_{1n}^* + g(x) \\ &= -\frac{1}{nh^p q} \frac{\sigma(x)}{f(x)} \sum_{k=1}^{q} \sum_{i=1}^{n} \frac{K_{ix} \eta_{i,k}^*}{f(c_k)} + g(x). \end{aligned} \tag{10}$$

Define the leave-one-out estimator $\hat{g}_{-i}$

$$\begin{aligned} \hat{g}_{-i} &= \hat{g}_{-i}(X_i) \\ &= -\frac{1}{nh^p q} \frac{\sigma(x)}{f(x)} \sum_{k=1}^{q} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{K_{ji} \eta_{j,k}^*}{f(c_k)} + g(x), \end{aligned} \tag{11}$$

where

$$\begin{aligned} K_{ji} &= W_{ji} L_{ji}, \\ W_{ji} &= W\left(\frac{X_j^c - X_i^c}{h}\right), \\ L_{ji} &= L_\lambda\left(X_j^d, X_i^d\right), \end{aligned}$$

and $\eta_{j,k}^*$ is defined similarly to $\eta_{i,k}^*$ in Section 2.

The loss function for the cross-validation exercise is given by

$$CV(h,\lambda) = n^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{g}_{-i}(X_i)\right]^2. \tag{12}$$

We let $\hat{h}$ and $\hat{\lambda}$ be the cross-validation solution of h and λ in Equation (12). Using Equation (1) and expanding the quadratic term in Equation (12) gives

$$\begin{aligned} CV(h,\lambda) &= n^{-1}\sum_{i=1}^{n}(g_i - \hat{g}_{-i})^2 + 2n^{-1}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i(g_i - \hat{g}_{-i}) + n^{-1}\sum_{i=1}^{n}\sigma(X_i)^2\varepsilon_i^2 \\ &= CV_1(h,\lambda) + n^{-1}\sum_{i=1}^{n}\sigma(X_i)^2\varepsilon_i^2, \end{aligned}$$

where

$$CV_1(h,\lambda) = n^{-1}\sum_{i=1}^{n}(g_i - \hat{g}_{-i})^2 + 2n^{-1}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i(g_i - \hat{g}_{-i}). \tag{13}$$

Here, the term $n^{-1}\sum_{i=1}^{n}\sigma(X_i)^2\varepsilon_i^2$ can be omitted in optimization as it is independent of both h and λ. Define

$$\begin{aligned} \gamma_i &= \sigma(X_i)/f(X_i), \\ \delta_i &= \sigma(X_i)^2/f(X_i). \end{aligned}$$

Substitute Equation (11) into the first term on the r.h.s. of Equation (13) and we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}(g_i - \hat{g}_{-i})^2 \\ &\quad = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{m=1}^{q}\frac{1}{f(c_k)f(c_m)}\frac{1}{n^3h^{2p}}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n}\gamma_i^2 K_{ji}K_{li}\eta_{j,k}^{*}\eta_{l,m}^{*}. \end{aligned} \tag{14}$$

Substitute Equation (11) into the second term on the r.h.s. of Equation (13) and we have

$$\begin{aligned} &\frac{2}{n}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i(g_i - \hat{g}_{-i}) \\ &\quad = \frac{2}{q}\sum_{k=1}^{q}\frac{1}{f(c_k)}\frac{1}{n^2h^p}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_i K_{ji}\eta_{j,k}^{*}\varepsilon_i. \end{aligned} \tag{15}$$

Hence, $CV_1(h,\lambda)$ can be written as

$$CV_1(h,\lambda) = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{m=1}^{q}\frac{1}{f(c_k)f(c_m)}S_{1,km} + \frac{2}{q}\sum_{k=1}^{q}\frac{1}{f(c_k)}S_{2,k}, \tag{16}$$

where

$$S_{1,km} = \frac{1}{n^3h^{2p}} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} \gamma_i^2 K_{ji} K_{li} \eta_{j,k}^* \eta_{l,m}^*,$$

$$S_{2,k} = \frac{1}{n^2h^p} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \delta_i K_{ji} \eta_{j,k}^* \varepsilon_i.$$

Lemmas 6 and 7 in Appendix A show that the leading term in cross-validation can be obtained in the following form:

$$CV_0 = A_1h^4 - A_2h^2\lambda + A_3\lambda^2 + A_4(nh^p)^{-1}, \tag{17}$$

where the coefficients A_1 to A_4 are constants defined in the proof of Theorem 3. Let the h_0 and λ_0 be the values of h and λ that minimize Equation (17). It is shown in the proof of Theorem 3 that h_0 and λ_0 are of order $O(n^{-1/(p+4)})$ and $O(n^{-2/(p+4)})$, respectively. The next theorem establishes the rates of convergence of $\left(\hat{h} - h_0\right)/h_0$ and $\left(\hat{\lambda} - \lambda_0\right)/\lambda_0$.

Theorem 3. *Under Assumptions 1 to 5, we have*

$$\left(\hat{h} - h_0\right)/h_0 = O_p(n^{-p/[2(p+4)]}) \text{ and } \left(\hat{\lambda} - \lambda_0\right)/ = O_p(n^{-1/2}).$$

Proof. See Appendix B. □

Remark 4. The results in Theorem 3 do not immediately prove that the use of $\hat{h}$ and $\hat{\lambda}$ lead to the asymptotic results in Equation (7), but it can be obtained in a two-step procedure. First, use h_0 and λ_0 to obtain the asymptotic result similar to Equation (7); second, use results in Theorem 3 to show that replacing h_0 and λ_0 with $\hat{h}$ and $\hat{\lambda}$ asymptotically gives the same distributional result. The proof will largely follow those in Theorem 1 and Theorem 2 and is not repeated here.

Remark 5. In collecting leading terms of $S_{1,km}$ in Lemma 6, the dimension parameter p appears in the exponent with the form of $-p$ and it does not appear in leading terms in $S_{2,k}$. This simplifies our analysis as compared to Theorem 2.2 in Racine and Li (2004) since there is no need to discuss separately the results when $p \leq 3$ and $p \geq 4$.

Remark 6. Similar to the results in Racine and Li (2004), the categorical regressor X_i^d also admits variables with a natural ordering such as self-reported financial status (low, middel, or high income group). The distance function can be modified to $d_{ix} = \sum_{t=1}^{r} |X_{t,i}^d - x_t^d|$ to model the density in $L_\lambda(X_i^d, x^d)$. Convergence results for $\hat{h}$ and $\hat{\lambda}$ can be obtained similarly to those in Theorem 3.

Remark 7. Based on Equations (30) and (32) in the Appendix, we know both $\hat{h}$ and $\hat{\lambda}$ are linear functions of B_1 and B_2, both of which are summations of zero mean random variables with finite variance. Hence, both B_1 and B_2 have an asymptotic normal distribution, which can be used to establish the asymptotic normality of $n^{p/[2(p+4)]}\left(\hat{h}-h\right)$ and $\sqrt{n}\left(\hat{\lambda}-\lambda\right)$.

Remark 8. A more sophisticated result can be derived by assuming different bandwidths for different components in both X_i^c and X_i^d. See Li and Racine (2004) for the rates of convergence of the smoothing parameters in this case. The results in Theorem 3 can also be extended to this case.

Remark 9. Kai, Li, and Zou (2010) estimate the optimal bandwidth based on the relation between optimal CQR bandwidth and optimal local least squares bandwidth.They also propose a 'pilot' selector method to calculate optimal CQR bandwidth. Results in our paper on using data-driven bandwidth selection to estimate CQR with mixed regressors complements the study in Kai, Li, and Zou (2010).

4 Monte Carlo Experiments

In this section, we use Monte Carlo simulations to illustrate the finite sample performance of local linear CQR with data-driven bandwidth selection. We compare our results to the local least square estimator in Racine and Li (2004) under several scenarios. Alhough we do not have a theoretical relative efficiency result due to the nonexistence of the closed form of optimal bandwidths based on MSEs, our Monte Carlo results document large efficiency gain of local CQR smoothers with cross-validated smoothing parameter to their local linear counterparts.

For both the CQR and local least squares estimate, cross-validation procedure is used to select the optimal bandwidth. The optimal bandwidth is then used to evaluate the mean function, $g(X)$, at 200 equally-spaced grid points for X^c on $[-1.5, 1.5]$ and 200 simulated $Bernoulli(0.5)$ random variables for X^d. We assess the performance of each estimator by calculating average squared errors (ASE) which is defined by

$$\text{ASE}(\hat{g}) = \frac{1}{200}\sum_{i=1}^{200}\left[\hat{g}(X_i) - g(X_i)\right]^2,$$

where $\hat{g}(X_i)$ is the estimate for $g(X_i)$. To compare the performance between the local least squares and CQR methods, we compute the ratio of average squared errors, $\text{RASE}(\hat{g}) = \text{ASE}(\hat{g}_{LS})/\text{ASE}(\hat{g}_{CQR})$, where $\hat{g}_{LS}$ and $\hat{g}_{CQR}$ are local least squares and CQR estimates. We simulate a sample of size 400 in each cross-validation exercise, select the optimal bandwidth, calculate ASE and RASE, and repeat the process 400 times. Tables 1 to 3 report the mean RASE over 400 replications and their associated standard deviations. In addition, we also report the mean ratios of optimal bandwidths selected in the CQR and LLS methods. A larger-than-one ratio for either h or λ means, on average, that the CQR uses larger bandwidth than LLS does.

Table 1: Simulation Results for DGP 1.

	RASE		Ratio of Bandwidths	
	Mean	Std. Dev.	h	λ
Standard Normal				
CQR_5	1.013	0.593	1.153	3.554
CQR_9	0.970	0.447	1.143	1.530
CQR_{19}	0.960	0.485	1.144	1.931
Laplace				
CQR_5	1.037	0.581	1.130	1.972
CQR_9	1.038	0.550	1.117	2.292
CQR_{19}	1.025	0.546	1.110	1.660
t_3				
CQR_5	1.078	0.469	1.172	2.441
CQR_9	1.054	0.4786	1.180	2.572
CQR_{19}	1.019	0.389	1.160	3.562
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,9)$				
CQR_5	1.074	0.557	1.077	2.111
CQR_9	1.037	0.552	1.105	2.042
CQR_{19}	1.058	0.576	1.108	2.129
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,100)$				
CQR_5	1.087	0.513	1.093	2.788
CQR_9	1.055	0.591	1.113	2.034
CQR_{19}	1.073	0.561	1.088	2.101

We consider the following three data-generating processes (DGPs) and the corresponding results are reported in Tables 1 to 3. The first DGP is given by

$$Y_i = \sin(2X_i^c) + 2\exp(-16(X_i^c)^2) + 0.5X_i^d + X_i^c X_i^d + 0.5\varepsilon_i,$$

where $X_i^c \sim N(0,1)$ and $X_i^d \sim \text{Bernoulli}(0.5)$. We also consider five error distributions for ε_i: standard normal, Laplace distribution with mean 0 and standard deviation 1, a t distribution with 3 degrees of freedom, and mixture normal distributions, $0.95N(0,1) + 0.05N(0,\sigma^2)$, with $\sigma = 3, 10$. We consider three choices of q in the CQR estimate: $q = 5, 9, 19$, referred to as CQR_5, CQR_9, CQR_{19} in the tables. This is exactly the same DGP as example 1 in Kai, Li, and Zou (2010).

The second DGP replaces $\sin(2X_i^c)$ in DGP 1 with X_i^c and everything else is the same as DGP1.

$$Y_i = X_i^c + 2\exp(-16(X_i^c)^2) + 0.5X_i^d + X_i^c X_i^d + 0.5\varepsilon_i.$$

In the third DGP, we let the categorical variable to appear in both the conditional mean and conditional variance, resembling the specification in Equation (1).

$$Y_i = X_i^c \sin(2\pi X_i^c) + 0.5X_i^d + X_i^c X_i^d + 10^{-1}(2 + \cos(2\pi X_i^c) + X_i^c X_i^d)\varepsilon_i,$$

where $X_i \sim U(0,1)$ and other specifications are the same as those in DGP 1. DGP 3 introduces a significant departure from DGPs 1 and 2 in that it incorporates not only heteroscedasticity but also categorical variables in the variance.

Table 2: Simulation Results for DGP 2.

	RASE		Ratio of Bandwidths	
	Mean	Std. Dev.	h	λ
Standard Normal				
CQR_5	1.074	0.689	1.133	2.363
CQR_9	1.035	0.622	1.131	2.265
CQR_{19}	1.049	0.616	1.109	1.939
Laplace				
CQR_5	1.151	0.949	1.117	2.391
CQR_9	1.100	0.806	1.106	2.011
CQR_{19}	1.055	0.700	1.119	2.545
t_3				
CQR_5	1.108	0.5171	1.120	2.697
CQR_9	1.085	0.520	1.135	2.653
CQR_{19}	1.041	0.433	1.129	3.554
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,9)$				
CQR_5	1.097	0.602	1.091	2.490
CQR_9	1.104	0.633	1.078	2.346
CQR_{19}	1.095	0.624	1.078	2.748
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,100)$				
CQR_5	1.139	0.609	1.085	2.878
CQR_9	1.122	0.610	1.077	2.234
CQR_{19}	1.109	0.599	1.082	2.840

In Table 1, under the standard normal error distribution, the mean RASE can be slightly below 1, indicating a small advantage of using LLS over CQR when $q = 9, 19$. For most of the cases in Table 1, we see that the CQR provides some small efficiency improvement over LLS. We also note that the selected bandwidth for CQR is larger than that in LLS, for both h and λ. In Table 2, we see that the CQR continues to provide some efficiency improvement over LLS. The improvement is large in the cases of Laplace and mixture normal error distributions.

Table 3 shows that there can be large efficiency gain in using CQR when the DGP exhibits heteroscedasticity, particularly when categorical variables also appear in the variance. It is also interesting to observe that, unlike results in Tables 1 and 2, the selected optimal bandwidths in LLS and CQR are very close for the continuous variable, although they still differ significantly for the categorical variable.

For each DGP under consideration, we observe that the CQR estimate provides better results as the error distribution becomes non-normal. In the three DGPs under study, the choice for λ for the CQR estimator seems to be always larger than that for the local least

Table 3: Simulation Results for DGP 3.

	RASE		Ratio of Bandwidths	
	Mean	Std. Dev.	h	λ
Standard Normal				
CQR_5	1.096	0.331	0.988	1.631
CQR_9	1.055	0.258	0.994	1.495
CQR_{19}	1.026	0.172	0.993	1.228
Laplace				
CQR_5	1.293	0.411	0.993	2.000
CQR_9	1.225	0.363	1.024	1.927
CQR_{19}	1.194	0.428	1.023	1.625
t_3				
CQR_5	1.555	0.643	0.993	1.849
CQR_9	1.422	0.528	0.991	2.251
CQR_{19}	1.293	0.382	0.985	1.480
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,9)$				
CQR_5	1.156	0.404	0.998	1.214
CQR_9	1.115	0.315	0.998	1.548
CQR_{19}	1.050	0.192	0.992	1.524
$0.95\mathcal{N}(0,1) + 0.05\mathcal{N}(0,100)$				
CQR_5	1.375	0.628	1.002	1.421
CQR_9	1.263	0.479	0.989	1.692
CQR_{19}	1.148	0.318	0.984	1.339

square estimator.

5 Conclusions

In this paper, we adopt the local CQR smoothing to nonparametric regression with mixed categorical and continuous regressors and propose cross-validation to select the optimal bandwidths for both types of regressors. We derive the asymptotic distribution of the local CQR estimator and the rates of convergence of bandwidths from cross-validation. Our simulation study shows efficiency gain of the proposed local CQR smoother under a variety of DGPs when compared to the local linear estimator. The efficiency gain may be large when categorical variables affect both the conditional mean and variance.

References

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Fan, Y., W. K. Härdle, W. Wang, and L. Zhu (2013): "Composite quantile regression for the single-index model," *working paper*.

Hall, P., Q. Li, and J. S. Racine (2007): "Nonparametric estimation of regression functions in the presence of irrelevant regressors," *The Review of Economics and Statistics*, 89(4), 784–789.

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

Härdle, W. (1990): *Applied nonparametric regression*. Cambridge university press.

Härdle, W., and O. Linton (1994): "Applied nonparametric methods," *Handbook of Econometrics*, 4, 2295–2339.

Henderson, D. J., and C. F. Parmeter (2015): *Applied nonparametric econometrics*. Cambridge University Press.

Horowitz, J. L. (2009): *Semiparametric and nonparametric methods in econometrics*. Springer.

Hsiao, C., Q. Li, and J. S. Racine (2007): "A consistent model specification test with mixed discrete and continuous data," *Journal of Econometrics*, 140(2), 802–826.

Ichimura, H., and P. E. Todd (2007): "Implementing nonparametric and semiparametric estimators," *Handbook of Econometrics*, 6, 5369–5468.

Jiang, R., Z.-G. Zhou, W.-M. Qian, and W.-Q. Shao (2012): "Single-index composite quantile regression," *Journal of the Korean Statistical Society*, 41(3), 323–332.

Kai, B., R. Li, and H. Zou (2010): "Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 49–69.

——— (2011): "New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models," *Annals of Statistics*, 39(1), 305.

Knight, K. (1998): "Limiting distributions for L_1 regression estimators under general conditions," *The Annals of Statistics*, 26(2), 755–770.

Koenker, R. (2004): "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91(1), 74–89.

Kong, E., and Y. Xia (2014): "An adaptive composite quantile approach to dimension reduction," *The Annals of Statistics*, 42(4), 1657–1688.

Li, D., and R. Li (2016): "Local composite quantile regression smoothing for Harris recurrent Markov processes," *Journal of Econometrics*, 194(1), 44–56.

Li, Q., E. Maasoumi, and J. S. Racine (2009): "A nonparametric test for equality of distributions with mixed categorical and continuous data," *Journal of Econometrics*, 148(2), 186–200.

Li, Q., and J. Racine (2003): "Nonparametric estimation of distributions with categorical and continuous data," *Journal of Multivariate Analysis*, 86(2), 266–292.

——— (2004): "Cross-validated local linear nonparametric regression," *Statistica Sinica*, 14(2), 485–512.

Li, Q., and J. S. Racine (2007): *Nonparametric econometrics: theory and practice*. Princeton University Press.

——— (2008): "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data," *Journal of Business & Economic Statistics*, 26(4), 423–434.

——— (2010): "Smooth varying-coefficient estimation and inference for qualitative and quantitative data," *Econometric Theory*, 26(6), 1607–1637.

Li, Q., J. S. Racine, and J. M. Wooldridge (2008): "Estimating average treatment effects with continuous and discrete covariates: The case of Swan-Ganz catheterization," *The American Economic Review*, 98(2), 357–362.

——— (2009): "Efficient estimation of average treatment effects with mixed categorical and continuous data," *Journal of Business & Economic Statistics*, 27(2), 206–223.

Ouyang, D., Q. Li, and J. S. Racine (2009): "Nonparametric estimation of regression functions with discrete regressors," *Econometric Theory*, 25(1), 1–42.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics*. Cambridge university press.

Pollard, D. (1991): "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7(2), 186–199.

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

Ruppert, D., and M. P. Wand (1994): "Multivariate locally weighted least squares regression," *The Annals of Statistics*, pp. 1346–1370.

Sun, Y. (2006): "A consistent nonparametric equality test of conditional quantile functions," *Econometric Theory*, 22(4), 614–632.

Zhao, Z., and Z. Xiao (2014): "Efficient regressions via optimally combining quantile information," *Econometric Theory*, 30(6), 1272–1314.

Appendix A Lemmas

Lemma 1. *Suppose Assumptions 1 to 5 hold, we have*

$$\frac{1}{nh^p}\sum_{i=1}^n K_{ix}/\sigma(X_i) = \frac{f(x)}{\sigma(x)} + O_p\Big(\lambda + h^2 + \frac{1}{\sqrt{nh^p}}\Big),$$
$$\frac{1}{nh^{p+1}}\sum_{i=1}^n K_{ix}(X_i^c - x^c)/\sigma(X_i) = h\mu_2\frac{\nabla_{x^c} f(x)}{\sigma(x)} + O_p\Big(\lambda h + h^3 + \frac{1}{\sqrt{nh^p}}\Big),$$
$$\frac{1}{nh^{p+2}}\sum_{i=1}^n K_{ix}(X_i^c - x^c)(X_i^c - x^c)^T/\sigma(X_i) = \mu_2\frac{f(x)}{\sigma(x)} + O_p\Big(\lambda + h^2 + \frac{1}{\sqrt{nh^p}}\Big).$$

Proof. The derivations are standard in nonparametric econometrics literature.

$$\mathbb{E}\Big[\frac{1}{nh^p}\sum_{i=1}^n K_{ix}/\sigma(X_i)\Big] = \frac{1}{h^p}\mathbb{E}(K_{ix}/\sigma(X_i))$$
$$= \frac{1}{h^p}\Big[\mathbb{E}(K_{ix}/\sigma(X_i)|d_{ix}=0)\cdot P(d_{ix}=0) + \mathbb{E}(K_{ix}/\sigma(X_i)|d_{ix}=1)\cdot P(d_{ix}=1)(1+O(\lambda))\Big]$$
$$= \frac{1}{h^p}\Big[\mathbb{E}(W_{ix}/\sigma(X_i)|d_{ix}=0)\cdot P(d_{ix}=0) + \lambda\mathbb{E}(W_{ix}/\sigma(X_i)|d_{ix}=1)\cdot P(d_{ix}=1)\cdot(1+O(\lambda))\Big]$$
$$= \int \frac{f(X_i^c, x^d)}{\sigma(X_i^c, x^d)} W\Big(\frac{X_i^c - x^c}{h}\Big) dX_i^c/h^p + \lambda \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \int \frac{f(X_i^c, \tilde{x}^d)}{\sigma(X_i^c, \tilde{x}^d)} W\Big(\frac{X_i^c - x^c}{h}\Big) dX_i^c/h^p \cdot (1+O(\lambda))$$
$$= \int \frac{f(x^c + hz, x^d)}{\sigma(x^c + hz, x^d)} W(z)dz + \lambda \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \int \frac{f(x^c + hz, \tilde{x}^d)}{\sigma(x^c + hz, \tilde{x}^d)} W(z)dz \cdot (1+O(\lambda))$$
$$= \frac{f(x)}{\sigma(x)} + O(h^2 + \lambda).$$

$$
\begin{aligned}
&Var\Big[\frac{1}{nh^p}\sum_{i=1}^n K_{ix}/\sigma(X_i)\Big] = \frac{1}{nh^{2p}}Var(K_{ix}/\sigma(X_i)) = \frac{1}{nh^{2p}}\Big[\mathbb{E}(K_{ix}^2/\sigma^2(X_i)) + O(h^{2p})\Big]\\
&\approx \frac{1}{nh^{2p}}\Big[\mathbb{E}(K_{ix}^2/\sigma^2(X_i)|d_{ix}=0)\cdot P(d_{ix}=0) + \mathbb{E}(K_{ix}^2/\sigma^2(X_i)|d_{ix}=1)\cdot P(d_{ix}=1) + O(h^{2p})\Big]\\
&= \frac{1}{nh^{2p}}\Big[\mathbb{E}(W_{ix}^2/\sigma^2(X_i)|d_{ix}=0)\cdot P(d_{ix}=0) + \lambda^2\mathbb{E}(W_{ix}^2/\sigma^2(X_i)|d_{ix}=1)\cdot P(d_{ix}=1) + O(h^{2p})\Big]\\
&= \frac{1}{nh^p}\Big[\int f(X_i^c, x^d)W^2\Big(\frac{X_i^c - x^c}{h}\Big)dX_i^c/h^p + O(\lambda^2 + h^{2p})\Big]\\
&= \frac{1}{nh^p}\nu_0\frac{f(x)}{\sigma^2(x)} + o\Big(\frac{1}{nh^p}\Big).
\end{aligned}
$$

Thus

$$
\frac{1}{nh^p}\sum_{i=1}^n K_{ix}/\sigma(X_i) = \frac{f(x)}{\sigma(x)} + O_p\Big(\lambda + h^2 + \frac{1}{\sqrt{nh^p}}\Big).
$$

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{nh^{p+1}}\sum_{i=1}^n \frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)\Big] = \frac{1}{h^{p+1}}\mathbb{E}\Big[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)\Big]\\
&\approx \frac{1}{h^{p+1}}\Big[\mathbb{E}\Big[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)|d_{ix}=0\Big]\cdot P(d_{ix}=0) + \mathbb{E}\Big[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)|d_{ix}=1\Big]\cdot P(d_{ix}=1)\Big]\\
&= \frac{1}{h^{p+1}}\Big[\mathbb{E}\Big[\frac{W_{ix}}{\sigma(X_i)}(X_i^c - x^c)|d_{ix}=0\Big]\cdot P(d_{ix}=0) + \lambda\mathbb{E}\Big[\frac{W_{ix}}{\sigma(X_i)}(X_i^c - x^c)|d_{ix}=1\Big]\cdot P(d_{ix}=1)\Big]\\
&= \frac{1}{h}\Big[\int \frac{f(X_i^c, x^d)}{\sigma(X_i^c, x^d)}W\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)dX_i^c/h^p + O(\lambda h^2)\Big]\\
&= \frac{1}{h}\int \frac{f(x^c + hz, x^d)}{\sigma(x^c + hz, x^d)}W(z)hzdz + O(\lambda h)\\
&= h\mu_2\frac{\nabla_{x^c} f(x)}{\sigma(x)} + O(\lambda h + h^3),
\end{aligned}
$$

and

$$
\begin{aligned}
&Var\Big[\frac{1}{nh^{p+1}}\sum_{i=1}^n \frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)\Big] = \frac{1}{nh^{2p+2}}Var[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)]\\
&= \frac{1}{nh^{2p+2}}\Big[\mathbb{E}\Big[\Big(\frac{K_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T] + O(h^{2p+4})\Big]\\
&\approx \frac{1}{nh^{2p+2}}\Big[\mathbb{E}\Big[\Big(\frac{W_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=0\Big]\cdot P(d_{ix}=0)\\
&\qquad + \lambda^2\mathbb{E}\Big[\Big(\frac{W_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=1)\cdot P(d_{ix}=1\Big] + O(h^{2p+4})\Big]\\
&= \frac{1}{nh^{p+2}}\Big[\int f(X_i^c, x^d)W^2\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c/h^p + O(\lambda^2h^2 + h^{2p+4})\Big]\\
&= \frac{1}{nh^p}\nu_2\frac{f(x)}{\sigma^2(x)} + o\Big(\frac{1}{nh^p}\Big).
\end{aligned}
$$

Thus

$$\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K_{ix}(X_i^c - x^c)/\sigma(X_i) = h\mu_2\frac{\nabla_{x^c}f(x)}{\sigma(x)} + O_p\Big(\lambda h + h^3 + \frac{1}{\sqrt{nh^p}}\Big).$$

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{nh^{p+2}}\sum_{i=1}^{n}\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T\Big] &= \frac{1}{h^{p+2}}\mathbb{E}[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T] \\
&= \frac{1}{h^{p+2}}\Big[\mathbb{E}[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=0]\cdot P(d_{ix}=0) \\
&\quad + \mathbb{E}[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=1]\cdot P(d_{ix}=1)(1+O(\lambda))\Big] \\
&= \frac{1}{h^{p+2}}\Big[\mathbb{E}[\frac{W_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=0]\cdot P(d_{ix}=0) \\
&\quad + \lambda\mathbb{E}[\frac{W_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=1]\cdot P(d_{ix}=1)(1+O(\lambda))\Big] \\
&= \frac{1}{h^2}\int\frac{f(X_i^c, x^d)}{\sigma(X_i^c, x^d)}W\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c/h^p \\
&\quad + \frac{1}{h^2}\lambda\sum_{\tilde{x}^d, d_{\tilde{x},x}=1}\int\frac{f(X_i^c, \tilde{x}^d)}{\sigma(X_i^c, \tilde{x}^d)}W\Big(\frac{X_i - x}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c/h^p \\
&= \mu_2\frac{f(x)}{\sigma(x)} + O(\lambda + h^2),
\end{aligned}$$

and

$$\begin{aligned}
&Var\Big[\frac{1}{nh^{p+2}}\sum_{i=1}^{n}\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T\Big] = \frac{1}{nh^{2p+4}}Var[\frac{K_{ix}}{\sigma(X_i)}(X_i^c - x^c)(X_i^c - x^c)^T] \\
&= \frac{1}{nh^{2p+4}}\Big[\mathbb{E}[\Big(\frac{K_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T\otimes(X_i^c - x^c)(X_i^c - x^c)^T] + O(h^{2p+4})\Big] \\
&= \frac{1}{nh^{2p+4}}\Big[\mathbb{E}(\Big(\frac{W_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T\otimes(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=0)\cdot P(d_{ix}=0) \\
&\quad + \lambda^2\mathbb{E}(\Big(\frac{W_{ix}}{\sigma(X_i)}\Big)^2(X_i^c - x^c)(X_i^c - x^c)^T\otimes(X_i^c - x^c)(X_i^c - x^c)^T|d_{ix}=1)\cdot P(d_{ix}=1)\Big] + O\Big(\frac{1}{n}\Big) \\
&= \frac{1}{nh^{p+4}}\Big[\int\frac{f(X_i^c, x^d)}{\sigma^2(X_i^c, x^d)}W^2\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T\otimes(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c/h^p \\
&\quad + \lambda^2\sum_{\tilde{x}^d, d_{\tilde{x},x}=1}\int f(X_i^c, \tilde{x}^d)W^2\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T\otimes(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c/h^p\Big] + O\Big(\frac{1}{n}\Big) \\
&= \frac{1}{nh^p}\nu_4\frac{f(x)}{\sigma^2(x)} + o\Big(\frac{1}{nh^p}\Big).
\end{aligned}$$

Finally, we have

$$\frac{1}{nh^{p+2}}\sum_{i=1}^{n}K_{ix}(X_i^c - x^c)(X_i^c - x^c)^T = \mu_2\frac{f(x)}{\sigma(x)} + O_p\Big(\lambda + h^2 + \frac{1}{\sqrt{nh^p}}\Big).$$

$\square$

Lemma 2. *Suppose Assumptions 1 to 5 hold, we have*

$$\frac{1}{nh^p}\sum_{i=1}^{n}K_{ix}^2 = \nu_0 f(x) + O_p\Big(\lambda^2 + h^2 + \frac{1}{\sqrt{nh^p}}\Big),$$
$$\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K_{ix}^2(X_i^c - x^c) = h\nu_2\nabla_{x^c}f(x) + O_p\Big(\lambda^2 h + h^3 + \frac{1}{\sqrt{nh^p}}\Big),$$
$$\frac{1}{nh^{p+2}}\sum_{i=1}^{n}K_{ix}^2(X_i^c - x^c)(X_i^c - x^c)^T = \nu_2 f(x) + O_p\Big(\lambda^2 + h^2 + \frac{1}{\sqrt{nh^p}}\Big).$$

Proof. To prove the first result, we have

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{nh^p}\sum_{i=1}^{n}K_{ix}^2\Big] = \frac{1}{h^p}\mathbb{E}\Big(K_{ix}^2\Big)\\
&\approx \frac{1}{h^p}\Big[\mathbb{E}\Big(K_{ix}^2|d_{ix}=0\Big)\cdot P(d_{ix}=0) + \mathbb{E}\Big(K_{ix}^2|d_{ix}=1\Big)\cdot P(d_{ix}=1)\Big]\\
&= \int f(X_i^c, x^d)W^2\Big(\frac{X_i^c - x^c}{h}\Big)dX_i^c/h^p + \lambda^2\sum_{\tilde{x}^d, d_{\tilde{x},x}=1}\int f(X_i^c, \tilde{x}^d)W^2\Big(\frac{X_i^c - x^c}{h}\Big)dX_i^c/h^p\\
&= \nu_0 f(x) + O(\lambda^2 + h^2),
\end{aligned}$$

and

$$\begin{aligned}
&Var\Big[\frac{1}{nh^p}\sum_{i=1}^{n}K_{ix}^2\Big] = \frac{1}{nh^{2p}}Var\Big(K_{ix}^2\Big) = \frac{1}{nh^{2p}}[\mathbb{E}(K_{ix}^4) + O(h^{2p})]\\
&\approx \frac{1}{nh^{2p}}\Big[\mathbb{E}(K_{ix}^4|d_{ix}=0)\cdot P(d_{ix}=0) + \mathbb{E}(K_{ix}^4|d_{ix}=1)\cdot P(d_{ix}=1) + O(h^{2p})\Big]\\
&= \frac{1}{nh^p}\Big[\int f(X_i^c, x^d)W^4\Big(\frac{X_i^c - x^c}{h}\Big)dX_i^c/h^p\\
&\quad + \lambda^4\sum_{\tilde{x}^d, d_{\tilde{x},x}=1}\int f(X_i^c, \tilde{x}^d)W^4\Big(\frac{X_i^c - x^c}{h}\Big)dX_i^c/h^p\Big] + O\Big(\frac{1}{n}\Big)\\
&= \frac{1}{nh^p}\int K^4(v)dvf(x) + o\Big(\frac{1}{nh^p}\Big).
\end{aligned}$$

Therefore,

$$\frac{1}{nh^p}\sum_{i=1}^{n}K_{ix}^2 = \nu_0 f(x) + O_p\Big(\lambda^2 + h^2 + \frac{1}{\sqrt{nh^p}}\Big).$$

Next, we have

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K_{ix}^2(X_i^c-x^c)\Big] = \frac{1}{h^{p+1}}\mathbb{E}[K_{ix}^2(X_i^c-x^c)]\\
&\approx \frac{1}{h^{p+1}}\Big[\mathbb{E}[W_{ix}^2(X_i^c-x^c)|d_{ix}=0]\cdot P(d_{ix}=0)+\lambda^2\mathbb{E}[W_{ix}^2(X_i^c-x^c)|d_{ix}=1]\cdot P(d_{ix}=1)\Big]\\
&=\frac{1}{h}\int f(X_i^c,x^d)W^2\Big(\frac{X_i^c-x^c}{h}\Big)(X_i^c-x^c)dX_i^c/h^p\\
&\quad+\frac{1}{h}\lambda^2\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\int f(X_i^c,\tilde{x}^d)W^2\Big(\frac{X_i^c-x^c}{h}\Big)(X_i^c-x^c)dX_i^c/h^p\\
&=h\nu_2\nabla_{x^c}f(x)+O(\lambda^2h+h^3),
\end{aligned}$$

and

$$\begin{aligned}
&Var\Big[\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K_{ix}^2(X_i^c-x^c)\Big] = \frac{1}{nh^{2p+2}}Var[K_{ix}^2(X_i^c-x^c)]\\
&=\frac{1}{nh^{2p+2}}[\mathbb{E}[K_{ix}^4(X_i^c-x^c)(X_i^c-x^c)^T]+O(h^{2p+4})]\\
&\approx\frac{1}{nh^{2p+2}}\Big[\mathbb{E}(W_{ix}^4(X_i^c-x^c)(X_i^c-x^c)^T|d_{ix}=0)\cdot P(d_{ix}=0)\\
&\quad+\lambda^4\mathbb{E}(W_{ix}^4(X_i^c-x^c)(X_i^c-x^c)^T|d_{ix}=1)\cdot P(d_{ix}=1)+O(h^{2p+4})\Big]\\
&=\frac{1}{nh^{p+2}}\Big[\int f(X_i^c,x^d)W^4\Big(\frac{X_i^c-x^c}{h}\Big)(X_i^c-x^c)(X_i^c-x^c)^TdX_i^c/h^p\\
&\quad+\lambda^4\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\int f(X_i^c,\tilde{x}^d)W^4\Big(\frac{X_i^c-x^c}{h}\Big)(X_i^c-x^c)(X_i^c-x^c)^TdX_i^c/h^p\Big]+O\Big(\frac{h^2}{n}\Big)\\
&=\frac{1}{nh^p}\int vv^TK^4(v)dvf(x)+o\Big(\frac{1}{nh^p}\Big).
\end{aligned}$$

Therefore,

$$\frac{1}{nh^{p+1}}\sum_{i=1}^{n}K_{ix}^2(X_i^c-x^c)=h\nu_2\nabla_{x^c}f(x)+O_p\Big(\lambda^2h+h^3+\frac{1}{\sqrt{nh^p}}\Big).$$

Finally, for the last result, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\frac{1}{nh^{p+2}}\sum_{i=1}^{n} K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T\Big] = \frac{1}{h^{p+2}}\mathbb{E}[K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T] \\
&\approx \frac{1}{h^{p+2}}\Big[\mathbb{E}[W_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T | d_{ix} = 0] \cdot P(d_{ix} = 0) \\
&\quad + \lambda^2 \mathbb{E}[W_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T | d_{ix} = 1] \cdot P(d_{ix} = 1)\Big] \\
&= \frac{1}{h^2}\int f(X_i^c, x^d) W^2\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c / h^p \\
&\quad + \frac{1}{h^2}\lambda^2 \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \int f(X_i^c, \tilde{x}^d) W^2\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T dX_i^c / h^p \\
&= \nu_2 f(x) + O(\lambda^2 + h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&Var\Big[\frac{1}{nh^{p+2}}\sum_{i=1}^{n} K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T\Big] = \frac{1}{nh^{2p+4}} Var[K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T] \\
&= \frac{1}{nh^{2p+4}}[\mathbb{E}[K_{ix}^4 (X_i^c - x^c)(X_i^c - x^c)^T \otimes (X_i^c - x^c)(X_i^c - x^c)^T] + O(h^{2p+4})] \\
&\approx \frac{1}{nh^{2p+4}}\Big[\mathbb{E}(W_{ix}^4 (X_i^c - x^c)(X_i^c - x^c)^T \otimes (X_i^c - x^c)(X_i^c - x^c)^T | d_{ix} = 0) \cdot P(d_{ix} = 0) \\
&\quad + \lambda^4 \mathbb{E}(W_{ix}^4 (X_i^c - x^c)(X_i^c - x^c)^T \otimes (X_i^c - x^c)(X_i^c - x^c)^T | d_{ix} = 1) \cdot P(d_{ix} = 1)\Big] + O\Big(\frac{1}{n}\Big) \\
&= \frac{1}{nh^{p+4}}\Big[\int f(X_i^c, x^d) W^4\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T \otimes (X_i^c - x^c)(X_i^c - x^c)^T dX_i^c / h^p \\
&\quad + \lambda^4 \sum_{\tilde{x}^d, d_{\tilde{x},x}=1} \int f(X_i^c, \tilde{x}^d) W^4\Big(\frac{X_i^c - x^c}{h}\Big)(X_i^c - x^c)(X_i^c - x^c)^T \otimes (X_i^c - x^c)(X_i^c - x^c)^T dX_i^c / h^p\Big] + O\Big(\frac{1}{n}\Big) \\
&= \frac{1}{nh^p}\int vv^T \otimes vv^T W^4(v) dv f(x) + o\Big(\frac{1}{nh^p}\Big).
\end{aligned}
$$

Thus

$$
\frac{1}{nh^{p+2}}\sum_{i=1}^{n} K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T = \nu_2 f(x) + O_p\Big(\lambda^2 + h^2 + \frac{1}{\sqrt{nh^p}}\Big).
$$

□

Lemma 3. *Suppose Assumptions 1 to 5 hold, we have*

$$
Var(W_n | X) \xrightarrow{P} f(x)\Sigma \equiv \Omega.
$$

Proof.

$$
\begin{aligned}
&Var(w_{1k}|X) = \mathbb{E}[w_{1k}^2|X] - 0 = \frac{1}{nh}\sum_{i=1}^{n} K_{ix}^2 \mathbb{E}[\eta_{i,k}^2|X] \\
&= \frac{1}{nh^p}\sum_{i=1}^{n} K_{ix}^2 \tau_{kk} \xrightarrow{p} \tau_{kk}\nu_0 f(x), k = 1, \cdots, q, \\
&Var(w_{21}|X) = \mathbb{E}[w_{21}w_{21}^T|X] - 0 = \frac{1}{nh^{p+2}}\sum_{i=1}^{n} \mathbb{E}\left(K_{ix}^2(X_i^c - x^c)(X_i^c - x^c)^T(\sum_{k=1}^{q}\eta_{i,k})^2\Big|X\right) \\
&= \frac{1}{nh^{p+2}}\sum_{i=1}^{n} K_{ix}^2(X_i^c - x^c)(X_i^c - x^c)^T \sum_{k,k'=1}^{q}\tau_{kk'} \xrightarrow{p} \nu_2 f(x)\sum_{k,k'=1}^{q}\tau_{kk'}, \\
&Cov(w_{1k}, w_{1k'}|X) = \mathbb{E}[w_{1k}\cdot w_{1k'}|X] - 0 = \frac{1}{nh^p}\sum_{i=1}^{n} K_{ix}^2 \mathbb{E}[\eta_{i,k}\eta_{i,k'}|X] \\
&= \frac{1}{nh^p}\sum_{i=1}^{n} K_{ix}^2 \tau_{kk'} \xrightarrow{p} \tau_{kk'}\nu_0 f(x), k \neq k', \\
&Cov(w_{1k}, w_{21}|X) = \mathbb{E}[w_{1k}\cdot w_{21}|X] = \frac{1}{nh^{p+1}}\sum_{i=1}^{n} K_{ix}^2(X_i^c - x^c)\sum_{k'=1}^{q}\mathbb{E}[\eta_{i,k}\eta_{i,k'}|X] \\
&= \frac{1}{nh^{p+1}}\sum_{i=1}^{n} K_{ix}^2(X_i^c - x^c)\sum_{k'=1}^{q}\tau_{kk'} \xrightarrow{p} 0.
\end{aligned}
$$

□

Based on the definition of d_{ix} in Section 2, let $d_{ji} = \sum_{t=1}^{r}\{X_{t,j}^d \neq X_{t,i}^d\}$.

Lemma 4. *Suppose Assumptions 1 to 5 hold. For all $i, j = 1, \cdots, n$ and $k = 1, \cdots, q$,*

$$
E\left(\eta_{j,k}^*|X_i, X_j\right) = \begin{cases} C_{1i,k}(X_j^c - X_i^c) + s.o. & \text{if } d_{ji} = 0, \\ C_{2i,k} + C_{3i,k}(X_j^c - X_i^c) + s.o. & \text{if } d_{ji} = 1, \end{cases}
$$

where $C_{1i,k}$, $C_{2i,k}$, and $C_{3i,k}$ are functions of X_i and are defined in the proof of this lemma.

Proof. Given the definition of $\eta_{j,k}^*$, we have

$$
\begin{aligned}
\mathbb{E}\left[\eta_{j,k}^*|X_i, X_j\right] &= \mathbb{E}\left[I\left(\varepsilon_j \le c_k - \frac{d_{j,k}}{\sigma(X_j)}\right) - \tau_k\right] \\
&= \int_{-\infty}^{c_k - \frac{d_{j,k}}{\sigma(X_j)}} f(v)dv - \tau_k \\
&= F\left(c_k - \frac{d_{j,k}}{\sigma(X_j)}\right) - F(c_k) \\
&= -\frac{d_{j,k}}{\sigma(X_j)} f(c_k) + o\left(\frac{d_{j,k}}{\sigma(X_j)}\right).
\end{aligned}
$$

Next, we analyze the term $-d_{j,k}/\sigma(X_j)$.

When $d_{ji} = 0$, for $d_{j,k}$, we have

$$
\begin{aligned}
d_{j,k} &= c_k \left[\sigma(X_j^c, X_i^d) - \sigma(X_i) \right] \\
&\quad + g(X_j^c, X_i^d) - g(X_i) - \beta(X_i)^T (X_j^c - X_i^c) \\
&= c_k \left[\sigma'(X_i)^T (X_j^c - X_i^c) + D_1 \right] + D_2,
\end{aligned}
$$

where the D_1 is defined as

$$
D_1 = O\left(\frac{1}{2} (X_j^c - X_i^c)^T \sigma''(X_i)(X_j^c - X_i^c) \right).
$$

Note that we will substitute $\mathbb{E}\left(\eta^*_{j,k} | X_i, X_j \right)$ into $\mathbb{E}\left(h^{-p} W_{ji} L_{ji} \eta^*_{j,k} | X_i \right)$ in Equation (23) and the order of D_1 becomes $h^2 v^2$ in integration, which is equivalent to an $O(h^2)$ term if we define $v = (X_j^c - X_i^c)/h$. Similarly, we have the $D_2 = O\left(\frac{1}{2} (X_j^c - X_i^c)^T \beta'(X_i)(X_j^c - X_i^c) \right) O\left((X_j^c - X_i^c)^2 \right)$ and it will become a small order term when combining with the kernel function W. Hereafter, we denote all such kind of terms as *s.o.* for notational simplicity.

For $1/\sigma(X_j)$, we have

$$
\frac{1}{\sigma(X_j)} = \frac{1}{\sigma(X_j^c, X_i^d)} = \frac{1}{\sigma(X_i)} - \frac{1}{\sigma^2(X_i)} \sigma'(X_i)(X_j^c - X_i^c) + s.o..
$$

In this case, it will be sufficient to keep the first term $\frac{1}{\sigma(X_i)}$ only and we have

$$
-\frac{d_{j,k}}{\sigma(X_j)} f(c_k) = C_{1i,k}(X_j^c - X_i^c) + s.o.,
$$

where

$$
C_{1i,k} = -f(c_k) c_k \sigma'(X_i)^T / \sigma(X_i).
$$

When $d_{ij} = 1$, we have

$$
\begin{aligned}
d_{j,k} &= c_k \left[\sigma(X_i^c, X_j^d) + \sigma'(X_i^c, X_j^d)^T (X_j^c - X_i^c) - \sigma(X_i) + s.o. \right] \\
&\quad + g(X_i^c, X_j^d) + \beta(X_i^c, X_j^d)^T (X_j^c - X_i^c) - g(X_i) - \beta(X_i)^T (X_j^c - X_i^c) \\
&= c_k \left[\sigma(X_i^c, X_j^d) - \sigma(X_i) \right] + g(X_i^c, X_j^d) - g(X_i) + c_k s.o. \\
&\quad + \left[\sigma'(X_i^c, X_j^d)^T + \beta(X_i^c, X_j^d)^T - \beta(X_i)^T \right] (X_j^c - X_i^c).
\end{aligned}
$$

Next,

$$
\begin{aligned}
\frac{1}{\sigma(X_j)} &= \frac{1}{\sigma(X_i^c, X_j^d)} - \frac{\sigma'(X_i^c, X_j^d)^T}{\sigma^2(X_i^c, X_j^d)} (X_j^c - X_i^c) + s.o. \\
&= \frac{1}{\sigma(X_i)} + \frac{\sigma(X_i) - \sigma(X_i^c, X_j^d)}{\sigma(X_i^c, X_j^d)\sigma(X_i)} - \frac{\sigma'(X_i^c, X_j^d)^T}{\sigma^2(X_i^c, X_j^d)} (X_j^c - X_i^c) + s.o. \qquad (18)
\end{aligned}
$$

It can be shown that the second term in Equation (18) is of order λ and the third term in Equation (18) is of order h in $\mathbb{E}\left(h^{-p}W_{ji}L_{ji}\eta^*_{j,k}|X_i\right)$, both of which can be omitted so that $1/\sigma(X_j) = 1/\sigma(X_i) + s.o.$ and

$$-\frac{d_{j,k}}{\sigma(X_j)}f(c_k) = C_{2i,k} + C_{3i,k}(X^c_j - X^c_i) + s.o., \tag{19}$$

where

$$\begin{aligned}
C_{2i,k} &= -f(c_k)\left[c_k\left(\sigma(X^c_i, X^d_j) - \sigma(X_i)\right) + g(X^c_i, X^d_j) - g(X_i)\right]/\sigma(X_i),\\
C_{3i,k} &= -f(c_k)\left[\sigma'(X^c_i, X^d_j)^T + \beta(X^c_i, X^d_j)^T - \beta(X_i)^T\right]/\sigma(X_i).
\end{aligned}$$

□

Lemma 5. *Suppose Assumptions 1 to 5 hold. For all $i, j = 1, \cdots, n$, and $k, m = 1, \cdots, q$,*

$$E\left(\eta^*_{j,k}\eta^*_{j,m}|X_i, X_j\right) = \begin{cases}(1-\tau_m)\tau_k + C_{4i,km}(X^c_j - X^c_i) + s.o. & \text{if } d_{ji} = 0,\\ C_{5i,km} + C_{6i,km}(X^c_j - X^c_i) + s.o. & \text{if } d_{ji} = 1,\end{cases}$$

where $C_{4i,km}$, $C_{5i,km}$, and $C_{6i,km}$ are functions of (X_i, τ_k, τ_m) and are defined in the proof of this lemma.

Proof. In taking expectations of $\eta^*_{j,k}\eta^*_{j,m}$, we assume $c_k - \frac{d_{j,k}}{\sigma(X_j)} \leq c_m - \frac{d_{j,m}}{\sigma(X_j)}$. The results for the case of $c_k - \frac{d_{j,k}}{\sigma(X_j)} \geq c_m - \frac{d_{j,m}}{\sigma(X_j)}$ will be the same.

$$\begin{aligned}
\mathbb{E}\left[\eta^*_{j,k}\eta^*_{j,m}|X_i, X_j\right] &= \mathbb{E}\left[1\left(\varepsilon_j \leq c_k - \frac{d_{j,k}}{\sigma(X_j)}\right)1\left(\varepsilon_j \leq c_m - \frac{d_{j,m}}{\sigma(X_j)}\right)|X_i, X_j\right]\\
&\quad - \tau_m\mathbb{E}\left[1\left(\varepsilon_j \leq c_k - \frac{d_{j,k}}{\sigma(X_j)}\right)|X_i, X_j\right] - \tau_k\mathbb{E}\left[1\left(\varepsilon_j \leq c_m - \frac{d_{j,m}}{\sigma(X_j)}\right)|X_i, X_j\right] + \tau_k\tau_m\\
&= (1-\tau_m)\mathbb{E}\left[1\left(\varepsilon_j \leq c_k - \frac{d_{j,k}}{\sigma(X_j)}\right)|X_i, X_j\right] - \tau_k\mathbb{E}\left[1\left(\varepsilon_j \leq c_m - \frac{d_{j,m}}{\sigma(X_j)}\right)|X_i, X_j\right] + \tau_k\tau_m.
\end{aligned} \tag{20}$$

Similar to Lemma 4, we analyze $\mathbb{E}\left[\eta^*_{j,k}\eta^*_{j,m}|X_i, X_j\right]$ under two cases: $d_{ji} = 0$ and $d_{ji} = 1$. Lemma 4 implies

$$E\left(\eta^*_{j,m}|X_i, X_j\right) = \begin{cases}C_{1i,m}(X^c_j - X^c_i) + s.o. & \text{if } d_{ji} = 0,\\ C_{2i,m} + C_{3i,m}(X^c_j - X^c_i) + s.o. & \text{if } d_{ji} = 1,\end{cases} \tag{21}$$

where $C_{1i,m}$, $C_{2i,m}$, and $C_{3i,m}$ are defined similarly to $C_{1i,k}$, $C_{2i,k}$, and $C_{3i,k}$ in Lemma 4.

Using both Lemma 4 and Equation (21), we obtain the following two results. When

$d_{ji} = 0$, Equation (20) becomes

$$\begin{aligned}
\mathbb{E}\left[\eta^*_{j,k}\eta^*_{j,m}|X_i, X_j\right] &= (1-\tau_m)\left(C_{1i,k}(X^c_j - X^c_i) + \tau_k\right) - \tau_k\left(C_{1i,m}(X^c_j - X^c_i) + \tau_m\right) + \tau_k\tau_m + s.o. \\
&= (1-\tau_m)\tau_k + C_{4i,km}(X^c_j - X^c_i) + s.o.,
\end{aligned}$$

where

$$C_{4i,km} = (1-\tau_m)C_{1i,k} - \tau_k C_{1i,m}.$$

When $d_{ji} = 1$, Equation (20) becomes

$$\begin{aligned}
\mathbb{E}\left[\eta^*_{j,k}\eta^*_{j,m}|X_i, X_j\right] &= (1-\tau_m)\left(C_{2i,k} + C_{3i,k}(X^c_j - X^c_i) + \tau_k\right) \\
&\quad - \tau_k\left(C_{2i,m} + C_{3i,m}(X^c_j - X^c_i) + \tau_m\right) + \tau_k\tau_m + s.o. \\
&= C_{5i,km} + C_{6i,km}(X^c_j - X^c_i) + s.o.,
\end{aligned}$$

where

$$\begin{aligned}
C_{5i,km} &= (1-\tau_m)\tau_k + (1-\tau_m)C_{2i,k} - \tau_k C_{2i,m} \\
C_{6i,km} &= (1-\tau_m)C_{3i,k} - \tau_k C_{3i,m}.
\end{aligned}$$

□

The following two lemmas use the U-statistics H-decomposition with variable kernels to calculate expectations of $S_{1,km}$ and $S_{2,k}$.

Lemma 6. *Suppose Assumptions 1 to 5 hold. For all $k, m = 1, \cdots, q$,*

$$\begin{aligned}
S_{1,km} &= A^*_{1,km}h^4 - A^*_{2,km}h^2\lambda + A^*_{3,km}\lambda^2 + A^*_{4,km}(nh)^{-1} \\
&\quad + \tilde{A}^*_{1,km}h^6 + \tilde{A}^*_{2,km}h^4\lambda + \tilde{A}^*_{3,km}h^2\lambda^2 + \tilde{A}^*_{4,km} + s.o.,
\end{aligned}$$

where the coefficients $A^*_{1,km}$, $A^*_{2,km}$, $A^*_{3,km}$ and $A^*_{4,km}$ are defined in the proof of this lemma.

Proof. Rewrite $S_{1,km}$ as

$$\begin{aligned}
S_{1,km} &= \frac{1}{n^3h^{2p}}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j\neq l}}^{n}\sum_{l=1}^{n}\gamma_i^2 W_{ji}L_{ji}W_{li}L_{li}\eta^*_{j,k}\eta^*_{l,m} + \frac{1}{n^3h^{2p}}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_i^2 W_{ji}^2 L_{ji}^2\eta^*_{j,k}\eta^*_{j,m} \\
&= S_{1,km,a} + S_{1,km,b}.
\end{aligned}$$

The first term $S_{1,km,a}$ can be written as

$$S_{1,km,a} = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j\neq l}}^{n}\sum_{l=1}^{n}H_{1a}(X^c_i, X^c_j, X^c_l),$$

where $H_{1a}(X_i^c, X_j^c, X_l^c)$ is a symmetrized version of $\gamma_i^2 W_{ji} L_{ji} W_{li} L_{li} \eta_{j,k}^* \eta_{l,m}^*$ and is given by

$$H_{1a}(X_i^c, X_j^c, X_l^c) = \frac{1}{3}(h^{-2p}\gamma_i^2 W_{ji} L_{ji} W_{li} L_{li} \eta_{j,k}^* \eta_{l,m}^* + h^{-2p}\gamma_j^2 W_{ij} L_{ij} W_{lj} L_{lj} \eta_{i,k}^* \eta_{l,m}^* + h^{-2p}\gamma_l^2 W_{jl} L_{jl} W_{il} L_{il} \eta_{i,k}^* \eta_{j,m}^*).$$

We first calculate the expectation of the following term.

$$\mathbb{E}\left[H_{1a}(X_i^c, X_j^c, X_l^c)\right] = \mathbb{E}\left\{\gamma_i^2 \mathbb{E}\left[h^{-p} W_{ji} L_{ji} \eta_{j,k}^* | X_i\right] \cdot \mathbb{E}\left[h^{-p} W_{li} L_{li} \eta_{l,m}^* | X_i\right]\right\}. \tag{22}$$

Calculation of $\mathbb{E}\left[h^{-1} W_{ji} L_{ji} \eta_{j,k}^* | X_i\right]$ will suffice.

$$\begin{aligned}
\mathbb{E}\left[h^{-p} W_{ji} L_{ji} \eta_{j,k}^* | X_i\right] &= \mathbb{E}\left[h^{-p} W_{ji} \eta_{j,k}^* | X_i, d_{ji} = 0\right] \cdot P\left(d_{ji} = 0 | X_i\right) \\
&\quad + \mathbb{E}\left[h^{-p} W_{ji} \eta_{j,k}^* | X_i, d_{ji} = 1\right] \cdot P\left(d_{ji} = 1 | X_i\right) \lambda \times (1 + O(\lambda)) \\
&= C_{1i,k} \int h^{-p} W_{ji} (X_j^c - X_i^c) f(X_j^c, X_i^d) dX_j^c \\
&\quad + \lambda C_{2i,k} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} \int h^{-p} W_{ji} f(X_j^c | \tilde{X}^d) dX_j^c \times P(d_{\tilde{x}i} = 1) \times (1 + O(\lambda)) \\
&\quad + \lambda C_{3i,k} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} \int h^{-p} W_{ji} (X_j^c - X_i^c) f(X_j^c | \tilde{X}^d) dX_j^c \times P(d_{\tilde{x}i} = 1) \times (1 + O(\lambda)) \\
&= C_{1i,k} \int W(v) hv \left[f(X_i) + f'(X_i) hv + \frac{1}{2} h^2 v^T f''(X_i) v + o(h^2)\right] dv \\
&\quad + \lambda C_{2i,k} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} \int W(v) \left[f(X_i^c, \tilde{X}^d) + hv f'(X_i^c, \tilde{X}^d) + o(h)\right] dv \times (1 + O(\lambda)) \\
&\quad + \lambda h C_{3i,k} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} \int v W(v) \left[f(X_i^c, \tilde{X}^d) + hv f'(X_i^c, \tilde{X}^d) + o(h)\right] dv \times (1 + O(\lambda)) \\
&= A_{1i,k} h^2 + O(h^4) - A_{2i,k} \lambda + O(\lambda h^2),
\end{aligned} \tag{23}$$

where

$$\begin{aligned}
A_{1i,k} &= C_{1i,k} \mu_2 f(X_i), \\
A_{2i,k} &= -C_{2i,k} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} f(X_i^c, \tilde{X}_i^d).
\end{aligned}$$

Similarly,

$$\mathbb{E}\left[h^{-p} W_{li} L_{li} \eta_{l,m}^* | X_i\right] = A_{1i,m} h^2 + O(h^4) - A_{2i,m} \lambda + O(\lambda h^2),$$

where,

$$\begin{aligned}
A_{1i,m} &= C_{1i,m} \mu_2 f(X_i), \\
A_{2i,m} &= -C_{2i,m} \sum_{\tilde{X}^d, d_{\tilde{x}i} = 1} f(X_i^c, \tilde{X}_i^d).
\end{aligned}$$

Substitute the results for both $\mathbb{E}\left[h^{-p} W_{ji} L_{ji} \eta_{j,k}^* | X_i\right]$ and $\mathbb{E}\left[h^{-p} W_{li} L_{li} \eta_{l,m}^* | X_i\right]$ into Equation (22) to get

$$\mathbb{E}\left[H_{1a}(X_i^c, X_j^c, X_l^c)\right] = A_{1,km}^* h^4 - A_{2,km}^* h^2\lambda + A_{3,km}^*\lambda^2 + \tilde{A}_{1,km}^* h^6 + \tilde{A}_{2,km}^* h^4\lambda + \tilde{A}_{3,km}^* h^2\lambda^2 + s.o.,$$

where

$$\begin{aligned}
A_{1,km}^* &= \mathbb{E}\left[\gamma_i^2 A_{1i,k} A_{1i,m}\right], \\
A_{2,km}^* &= \mathbb{E}\left[\gamma_i^2\left(A_{1i,k}A_{2i,m} + A_{1i,m}A_{2i,k}\right)\right], \\
A_{3,km}^* &= \mathbb{E}\left[A_{2i,k}A_{2i,m}\right].
\end{aligned}$$

The expressions for $\tilde{A}_{1,km}^*$, $\tilde{A}_{2,km}^*$, and $\tilde{A}_{3,km}^*$ are omitted as they are associated with terms of smaller order and are not used in the proof.

Next, consider $S_{1,km,b}$.

$$S_{1,km,b} = \frac{1}{n}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} H_{1b}(X_i^c, X_j^c),$$

where H_{1b} is a symmetricized version of $h^{-2p}\gamma_i^2 W_{ji}^2 L_{ji}^2 \eta_{j,k}^*\eta_{j,m}^*$ and it takes the following form:

$$H_{1b}(X_i^c, X_j^c) = \frac{1}{2}\left(h^{-2p}\gamma_i^2 W_{ji}^2 L_{ji}^2\eta_{j,k}^*\eta_{j,m}^* + h^{-2p}\gamma_j^2 W_{ij}^2 L_{ij}^2\eta_{i,k}^*\eta_{i,m}^*\right).$$

We utilize Lemma 5 to calculate the expectation of $\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)\right]$.

$$\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)\right] = \mathbb{E}\left[\gamma_i^2\mathbb{E}\left[h^{-2p}W_{ji}^2L_{ji}^2\eta_{j,k}^*\eta_{j,m}^*|X_i\right]\right]$$

and

$$\begin{aligned}
&\mathbb{E}\left[h^{-2p}W_{ji}^2L_{ji}^2\eta_{j,k}^*\eta_{j,m}^*|X_i\right] = \mathbb{E}\left[h^{-2p}W_{ji}^2\eta_{j,k}^*\eta_{j,m}^*|X_i, d_{ji}=0\right]\cdot P(d_{ji}=0|X_i) \\
&\quad + \mathbb{E}\left[h^{-2p}W_{ji}^2\eta_{j,k}^*\eta_{j,m}^*|X_i, d_{ji}=1\right]\cdot P(d_{ji}=1|X_i)\lambda^2\times(1+O(\lambda^2)) \\
&= (1-\tau_m)\tau_k h^{-p}\int h^{-p}W_{ji}^2 f(X_j^c, X_j^d)dX_j^c \\
&\quad + h^{-p}C_{4i,km}\int h^{-p}W_{ji}^2(X_j^c - X_i^c)f(X_j^c, X_j^d)dX_j^c \\
&\quad + \lambda^2\sum_{\tilde{X}^d, d_{\tilde{x}i}}\int h^{-p}W_{ji}^2[C_{5i,km} + C_{6i,km}(X_j^c - X_i^c)]f(X_j^c|\tilde{X}^d)dX_j^c\times P(d_{\tilde{x}i}=1)\times(1+O(\lambda^2)) \\
&= (1-\tau_m)\tau_k h^{-p}\int W^2(v)\left[f(X_i) + f'(X_i)hv + O(h^2)\right]dv \\
&\quad + C_{4i,km}\int W^2(v)v\left[f(X_i) + f'(X_i)hv + O(h^2)\right]dv + O(\lambda^2) \\
&= (1-\tau_m)\tau_k h^{-p}f(X_i)\int W^2(v)dv + s.o.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)\right] = A_{4,km}^* h^{-p},$$

where

$$A_{4,km}^* = (1-\tau_m)\tau_k v_0 \mathbb{E}\left[\gamma_i^2 f(X_i)\right].$$

Similarly, we can show $\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)|X_i\right] = O(h^{-p})$. Hence, by H-decomposition, we have

$$\begin{aligned}
S_{1,km,b} &= n^{-1}\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)\right] \\
&\quad + 2n^{-1}\sum_{i=1}^{n}\left\{\mathbb{E}\left[H_{1b}(X_i^c, X_j^c)|X_i\right] - \mathbb{E}\left[H_{1b}(X_i^c, X_j^c)\right]\right\} + s.o. \\
&= A_{4,km}^*(nh^p)^{-1} + n^{-1/2}O((nh^p)^{-1}).
\end{aligned}$$

Combining the results for $S_{1,km,a}$ and $S_{1,km,b}$, we have

$$S_{1,km} = A_{1,km}^* h^4 - A_{2,km}^* h^2\lambda + A_{3,km}^*\lambda^2 + A_{4,km}^*(nh^p)^{-1} + \tilde{A}_{1,km}^* h^6 + \tilde{A}_{2,km}^* h^4\lambda + \tilde{A}_{3,km}^* h^2\lambda^2 + s.o.$$

□

Lemma 7. *For all $k = 1, \cdots, q$,*

$$S_{2,k} = B_{1,k}^*\frac{h^2}{\sqrt{n}} + B_{2,k}^*\frac{\lambda}{\sqrt{n}} + s.o.,$$

where the coefficients $B_{1,k}^*$ and $B_{2,k}^*$ are defined in the proof of this lemma.

Proof. Write

$$S_{2,k} = n^{-2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} H_2(X_i^c, X_j^c),$$

where

$$H_2(X_i^c, X_j^c) = \frac{1}{2}\left(h^{-p}\delta_i W_{ji}L_{ji}\eta_{j,k}^*\varepsilon_i + h^{-p}\delta_j W_{ij}L_{ij}\eta_{i,k}^*\varepsilon_j\right).$$

Using Equation (23) and the independence between ε_i and $\eta_{j,k}^*$, we have

$$\begin{aligned}
\mathbb{E}\left[\delta_i h^{-p} W_{ji}L_{ji}\eta_{j,k}^*\varepsilon_i|X_i\right] &= \delta_i\varepsilon_i\mathbb{E}\left[h^{-p}W_{ji}L_{ji}\eta_{j,k}^*|X_i\right] \\
&= \delta_i\varepsilon_i\left(A_{1i,k}h^2 - A_{2i,k}\lambda + O(h^4) + O(\lambda h^2)\right) \\
&= B_{11i,k}h^2 + B_{12i,k}\lambda + O(h^4) + O(\lambda h^2),
\end{aligned}$$

where $B_{11i,k} = \delta_i\varepsilon_i A_{1i,k}$ and $B_{12i,k} = \delta_i\varepsilon_i A_{2i,k}$. Using H-decomposition similar to that for

$S_{1,km,b}$ in Lemma 6 and the fact that $\mathbb{E}\left[H_2(X_i^c, X_j^c)\right] = 0$, we have

$$\begin{aligned} S_{2,k} &= \frac{2}{n}\sum_{i=1}^{n} B_{11,k}h^2 + \frac{2}{n}\sum_{i=1}^{n} B_{12,k}\lambda + s.o. \\ &= B_{1,k}^{*}\frac{h^2}{\sqrt{n}} + B_{2,k}^{*}\frac{\lambda}{\sqrt{n}} + s.o., \end{aligned}$$

where

$$\begin{aligned} B_{1,k}^{*} &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n} B_{11i,k} \\ B_{2,k}^{*} &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n} B_{12i,k} \end{aligned}$$

and both $B_{1,k}^{*}$ and $B_{2,k}^{*}$ are $O_p(1)$. □

Appendix B Proofs

Proof of Proposition 1. To apply the identity (Knight (1998); Koenker (2004); Kai, Li, and Zou (2010)), we write $Y_i - a_k - b(X_i^c - x^c) = \sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} - \Delta_{i,k}$. Equation (3) is equivalent to minimize (by Knight identity)

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n}\left\{K_{ix}\sum_{k=1}^{q}\left[\rho_{\tau_k}\Big(\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} - \Delta_{i,k}\Big) - \rho_{\tau_k}\Big(\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k}\Big)\right]\right\}, \\ &= \sum_{i=1}^{n}\Bigg\{K_{ix}\sum_{k=1}^{q}\Big[\Delta_{i,k}\big[1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le 0\} - \tau_k\big] \\ &\qquad + \int_0^{\Delta_{i,k}}\big[1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le w\} - 1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le 0\}\big]dw\Big]\Bigg\}, \\ &= \sum_{i=1}^{n}\left(K_{ix}\sum_{k=1}^{q}\left[\left(\frac{u_k}{\sqrt{nh^p}} + \frac{v^T(X_i^c - x^c)}{\sqrt{nh^{p+2}}}\right)\big[1\{\sigma(X_i)(\varepsilon_i - c_k) + d_{i,k} \le 0\} - \tau_k\big]\right]\right) \\ &\qquad + \sum_{k=1}^{q} B_{n,k}(\theta), \\ &= \sum_{k=1}^{q} u_k\left(\sum_{i=1}^{n}\frac{K_{ix}\eta_{i,k}^{*}}{\sqrt{nh^p}}\right) + v^T\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_{ix}(X_i^c - x^c)\eta_{i,k}^{*}}{\sqrt{nh^{p+2}}}\right) + \sum_{k=1}^{q} B_{n,k}(\theta). \end{aligned}$$

□

Proof of Proposition 2. Write $L_n(\theta)$ as

$$L_n(\theta) = \sum_{k=1}^{q} u_k\left(\sum_{i=1}^{n}\frac{K_{ix}\eta_{i,k}^{*}}{\sqrt{nh^p}}\right) + v^T\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_{ix}(X_i^c - x^c)\eta_{i,k}^{*}}{\sqrt{nh^{p+2}}}\right) + \sum_{k=1}^{q}\mathbb{E}_{\varepsilon}[B_{n,k}(\theta)|X] + \sum_{k=1}^{q} R_{n,k}(\theta),$$

where $R_{n,k}(\theta) = B_{n,k}(\theta) - \mathbb{E}_\varepsilon[B_{n,k}(\theta)|X]$. Using $F_\varepsilon(c_k + w) - F_\varepsilon(c_k) = wf(c_k) + o(w)$, then

$$\begin{aligned}
\sum_{k=1}^{q} \mathbb{E}_\varepsilon[B_{n,k}(\theta)|X] &= \sum_{k=1}^{q}\sum_{i=1}^{n} K_{ix} \int_0^{\Delta_{i,k}} \mathbb{E}_\varepsilon \left[1\{\varepsilon_i \le c_k + \frac{w - d_{i,k}}{\sigma(X_i)}\} - 1\{\varepsilon_i \le c_k - \frac{d_{i,k}}{\sigma(X_i)}\} \Big| X \right] dw \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n} \Big(K_{ix} \int_0^{\Delta_{i,k}} \left[\frac{w}{\sigma(X_i)} f_\varepsilon\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big) + o(w) \right] dw \Big) \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n} \left[K_{ix} \frac{\Delta_{i,k}^2}{2\sigma(X_i)} f_\varepsilon\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big) \right] + o_p(1) \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n} \Big[K_{ix} \frac{\Delta_{i,k}^2}{2\sigma(X_i)} f_\varepsilon(c_k) \Big] + o_p(1) \\
&= \frac{1}{2}\theta^T S_n \theta + o_p(1).
\end{aligned}$$

We now prove that $R_{n,k}(\theta) = o_p(1)$. It is sufficient to show that $Var_\varepsilon[B_{n,k}(\theta)|X] = o_p(1)$. In fact,

$$\begin{aligned}
Var_\varepsilon[B_{n,k}(\theta)|X] &= \sum_{i=1}^{n} Var_\varepsilon \Big[\Big(K_{ix} \int_0^{\Delta_{i,k}} \Big(1\Big\{\varepsilon_i \le c_k - \frac{d_{i,k}}{\sigma(X_i)} + \frac{w}{\sigma(X_i)}\Big\} - 1\Big\{\varepsilon_i \le c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big\}\Big) dw \Big) \Big| X \Big] \\
&\le \sum_{i=1}^{n} \mathbb{E}_\varepsilon \Big[\Big(K_{ix} \int_0^{\Delta_{i,k}} \Big(1\Big\{\varepsilon_i \le c_k - \frac{d_{i,k}}{\sigma(X_i)} + \frac{w}{\sigma(X_i)}\Big\} - 1\Big\{\varepsilon_i \le c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big\}\Big) dw \Big)^2 \Big| X \Big] \\
&\le \sum_{i=1}^{n} K_{ix}^2 \int_0^{|\Delta_{i,k}|} \int_0^{|\Delta_{i,k}|} \Big[F\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)} + \frac{|\Delta_{i,k}|}{\sigma(X_i)}\Big) - F\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big) \Big] dw_1 dw_2 \\
&= o\Big(\sum_{i=1}^{n} K_{ix}^2 \Delta_{i,k}^2 \Big) = o_p(1).
\end{aligned}$$

□

Proof of Theorem 1. From Lemma 2, we have

$$S_n \xrightarrow{P} \frac{f(x)}{\sigma(x)} S = \frac{f(x)}{\sigma(x)} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

Together with Propositions 1 and 2, we have

$$L_n(\theta) = \frac{1}{2}\frac{f(x)}{\sigma(x)} \theta' S \theta + (W_n^*)'\theta + o_p(1).$$

Since the convex function $L_n(\theta) - (W_n^*)'\theta$ converges in probability to the convex function

$$\frac{1}{2}\frac{f(x)}{\sigma(x)} \theta' S \theta,$$

it follows from the convexity lemma (Pollard (1991)) that, for any compact set Θ, the quadratic approximation to $L_n(\theta)$ holds uniformly for θ in any compact set, which leads to

$$\hat{\theta}_n = -\frac{\sigma(x)}{f(x)} S^{-1} W_n^* + o_p(1). \tag{24}$$

By the Cramer-Wald theorem, it is easy to see that the central limit theorem for $W_n|X$ holds:

$$\frac{W_n|X - \mathbb{E}[W_n|X]}{\sqrt{Var(W_n|X)}} \xrightarrow{d} MVN(0, I_{q+p}).$$

Note that

$$\begin{aligned}
Cov(\eta_{i,k}, \eta_{i,k'}) &= Cov(1\{\varepsilon_i \le c_k\} - \tau_k, 1\{\varepsilon_i \le c_{k'}\} - \tau_{k'}) = Cov(1\{\varepsilon_i \le c_k\}, 1\{\varepsilon_i \le c_{k'}\}) \\
&= \mathbb{E}[1\{\varepsilon_i \le c_k\} \times 1\{\varepsilon_i \le c_{k'}\}] - \tau_k \cdot \tau_{k'} = \tau_k \wedge \tau_{k'} - \tau_k \cdot \tau_{k'} \\
&= \tau_{kk'},
\end{aligned}$$

and

$$\begin{aligned}
Cov(\eta_{i,k}, \eta_{j,k'}) &= Cov(1\{\varepsilon_i \le c_k\} - \tau_k, 1\{\varepsilon_j \le c_{k'}\} - \tau_{k'}) = Cov(1\{\varepsilon_i \le c_k\}, 1\{\varepsilon_j \le c_{k'}\}) \\
&= \mathbb{E}[1\{\varepsilon_i \le c_k\} \times 1\{\varepsilon_j \le c_{k'}\}] - \tau_k \cdot \tau_{k'} = \tau_k \cdot \tau_{k'} - \tau_k \cdot \tau_{k'} \\
&= 0, \text{ if } i \ne j.
\end{aligned}$$

Further, for $k = 1, \cdots, q$,

$$\begin{aligned}
\mathbb{E}[w_{1k}|X] &= \frac{1}{\sqrt{nh^p}} \sum_{i=1}^n \mathbb{E}[K_{ix}\eta_{i,k}|X] = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_{ix}\mathbb{E}[\eta_{i,k}|X] = 0, \\
\mathbb{E}[w_{21}|X] &= \frac{1}{\sqrt{nh^{p+2}}} \sum_{k=1}^q \sum_{i=1}^n \mathbb{E}[K_{ix}(X_i^c - x^c)\eta_{i,k}|X] = \frac{1}{\sqrt{nh^{p+2}}} \sum_{k=1}^q \sum_{i=1}^n K_{ix}(X_i^c - x^c)\mathbb{E}[\eta_{i,k}|X] = 0,
\end{aligned}$$

$Var(W_n|X) \xrightarrow{P} f(x)\Sigma.$

Therefore

$$W_n|X \xrightarrow{d} MVN\Big(0, f(x)\Sigma\Big).$$

Moreover, we have

$$\begin{aligned}
Var(w_{1k}^* - w_{1k}|X) &= \frac{1}{nh^p} \sum_{i=1}^n K_{ix}^2 Var(\eta_{i,k}^* - \eta_{i,k}|X) \le \frac{1}{nh} \sum_{i=1}^n K_{ix}^2 \Big[F\Big(c_k + \frac{|d_{i,k}|}{\sigma(X_i)}\Big) - F(c_k)\Big] = o_p(1), \\
Var(w_{21}^* - w_{21}|X) &= \frac{1}{nh^{p+2}} \sum_{i=1}^n K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T Var[\sum_{k=1}^q (\eta_{i,k}^* - \eta_{i,k})|X] \\
&\le \frac{1}{nh^{p+2}} \sum_{i=1}^n K_{ix}^2 (X_i^c - x^c)(X_i^c - x^c)^T \max_k \Big[F\Big(c_k + \frac{|d_{i,k}|}{\sigma(X_i)}\Big) - F(c_k)\Big] = o_p(1).
\end{aligned}$$

Thus $Var(W_n^* - W_n|X) = o_p(1)$. So by Slutsky's theorem, conditioning on X, we have $W_n^*|X - \mathbb{E}(W_n^*|X) \xrightarrow{d} MVN\Big(0, f(x)\Sigma\Big)$.

Therefore

$$\hat{\theta}_n + \frac{\sigma(x)}{f(x)} S^{-1}\mathbb{E}(W_n^*|X) \xrightarrow{d} MVN\Big(0, \frac{\sigma^2(x)}{f(x)} S^{-1}\Sigma S^{-1}\Big).$$

$\square$

Proof of Theorem 2. Denote $e_{q\times 1}$ the vector with q 1's. S^{-1} is a diagonal matrix

$$S^{-1} = \begin{pmatrix} f(c_1) & 0 & \cdots & 0 & 0 \\ 0 & f(c_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & f(c_q) & 0 \\ 0 & 0 & \cdots & 0 & \mu_2 \sum_{k=1}^{q} f(c_k) \end{pmatrix}^{-1}.$$

By the definition of θ and u_k, we have

$$a_k = \frac{u_k}{\sqrt{nh^p}} + g(x) + \sigma(x)c_k.$$

$$\begin{aligned} \mathbb{E}(\hat{g}(x)|X) &= g(x) + \frac{1}{q}\sum_{k=1}^{q}\sigma(x)c_k + \frac{1}{q\sqrt{nh^p}}\sum_{k=1}^{q}\mathbb{E}(\hat{u}_k|X) \\ &= g(x) + \frac{1}{q}\sum_{k=1}^{q}\sigma(x)c_k - \frac{\sigma(x)}{q\sqrt{nh^p}f(x)} e_{q\times 1}(S^{-1})_{11}\mathbb{E}(W_{1n}^*|X), \end{aligned}$$

where W_{1n}^* is the first q elements of W_n^* and $(S^{-1})_{11}$ is the upper-left $q\times q$ corner matrix of S^{-1}. Therefore,

$$\begin{aligned} Bias(\hat{g}(x)|X) &= \frac{1}{q}\sum_{k=1}^{q}\sigma(x)c_k - \frac{\sigma(x)}{q\sqrt{nh^p}f(x)} e_{q\times 1}(S^{-1})_{11}\mathbb{E}(W_{1n}^*|X) \\ &= \frac{1}{q}\sum_{k=1}^{q}\sigma(x)c_k - \frac{1}{qnh^p}\frac{\sigma(x)}{f(x)}\sum_{i=1}^{n}K_{ix}\sum_{k=1}^{q}\frac{1}{f(c_k)}\Big[F\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big) - F(c_k)\Big]. \end{aligned}$$

Note that the error is symmetric; thus $\sum_{k=1}^{q} c_k = 0$ and, furthermore, it is easy to check that

$$\begin{aligned} \frac{1}{q}\sum_{k=1}^{q}\frac{1}{f(c_k)}\Big[F\Big(c_k - \frac{d_{i,k}}{\sigma(X_i)}\Big) - F(c_k)\Big] &= \frac{1}{q}\sum_{k=1}^{q}\frac{1}{f(c_k)}f(c_k)\times\Big[-\frac{d_{i,k}}{\sigma(X_i)} + o_p\Big(\frac{d_{i,k}}{\sigma(X_i)}\Big)\Big] \\ &= \frac{1}{q}\sum_{k=1}^{q}\Big[-\frac{c_k(\sigma(X_i)-\sigma(x,z)) + r_i}{\sigma(X_i)}\Big](1+o_p(1)) \\ &= -\frac{r_i}{\sigma(X_i)}(1+o_p(1)). \end{aligned}$$

Therefore,

$$\begin{aligned}
Bias(\hat{g}(x)|X) &= -\frac{1}{qnh^p}\frac{\sigma(x)}{f(x)}\sum_{i=1}^{n}K_{ix}\sum_{k=1}^{q}\frac{1}{f(c_k)}\Big[F\Big(c_k-\frac{d_{i,k}}{\sigma(X_i)}\Big)-F(c_k)\Big]\\
&= -\frac{1}{nh^p}\frac{\sigma(x)}{f(x)}\sum_{i=1}^{n}K_{ix}\times\Big[-\frac{r_i}{\sigma(X_i)}(1+o_p(1))\Big]\\
&= -\frac{1}{nh^p}\frac{\sigma(x)}{f(x)}\sum_{i=1}^{n}K_{ix}\times\Big[-\frac{g(X_i)-g(x)-\beta(x)(X_i^c-x^c)}{\sigma(X_i)}\Big]\\
&= \Big[h^2\frac{tr[\beta'(x)]\mu_2}{2}+\lambda\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\frac{\sigma(x)f(x^c,\tilde{x}^d)[g(x^c,\tilde{x}^d)-g(x)]}{f(x)\sigma(x^c,\tilde{x}^d)}\Big](1+o_p(1)),
\end{aligned}$$

when using

$$\begin{aligned}
&\mathbb{E}\Big[\frac{1}{h^p}K_{ix}r_i/\sigma(X_i)\Big]\approx\frac{1}{h^p}\mathbb{E}[K_{ix}r_i/\sigma(X_i)|d_{ix}=0]\cdot P(d_{ix}=0)+\frac{1}{h^p}\mathbb{E}[K_{ix}r_i/\sigma(X_i)|d_{ix}=1]\cdot P(d_{ix}=1)\\
&=\frac{1}{h^p}\mathbb{E}[W_{ix}r_i/\sigma(X_i)|d_{ix}=0]\cdot P(d_{ix}=0)+\frac{\lambda}{h^p}\mathbb{E}[W_{ix}r_i/\sigma(X_i)|d_{ix}=1]\cdot P(d_{ix}=1)\\
&=\int f(X_i^c,x^d)W\Big(\frac{X_i^c-x^c}{h}\Big)[g(X_i^c,x^d)-g(x)-\beta(x)(X_i^c-x^c)]/\sigma(X_i^c,x^d)dX_i^c/h^p\\
&+\lambda\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\int f(X_i^c,\tilde{x}^d)W\Big(\frac{X_i^c-x^c}{h}\Big)[g(X_i^c,\tilde{x}^d)-g(x)-\beta(x)(X_i^c-x^c)]/\sigma(X_i^c,\tilde{x}^d)dX_i^c/h^p\\
&=h^2\frac{f(x)tr[\beta'(x)]\mu_2}{2\sigma(x)}+O(h^4)+\lambda\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\frac{f(x^c,\tilde{x}^d)[g(x^c,\tilde{x}^d)-g(x)]}{\sigma(x^c,\tilde{x}^d)}+O(\lambda^2)\\
&=h^2\frac{f(x)tr[\beta'(x)]\mu_2}{2\sigma(x)}+\lambda\sum_{\tilde{x}^d,d_{\tilde{x},x}=1}\frac{f(x^c,\tilde{x}^d)[g(x^c,\tilde{x}^d)-g(x)]}{\sigma(x^c,\tilde{x}^d)}+o(h^2+\lambda).
\end{aligned}$$

Furthermore, the conditional variable of $\hat{g}(x)$ is

$$\begin{aligned}
Var[\hat{g}(x)|X] &= \frac{1}{nh^p}\frac{\sigma^2(x)}{f(x)}\frac{1}{q^2}e'_{q\times1}(S^{-1}\Sigma S^{-1})_{11}e_{q\times1}+o_p(\frac{1}{nh^p})\\
&= \frac{1}{nh^p}\frac{\sigma^2(x)}{f(x)}\frac{1}{q^2}\sum_{k=1}^{q}\sum_{k'=1}^{q}\frac{\nu_0\tau_{kk'}}{f(c_k)f(c_{k'})}+o_p(\frac{1}{nh})\\
&= \frac{1}{nh^p}\frac{\sigma^2(x)}{f(x)}\nu_0R_1(q)+o_p(\frac{1}{nh^p}).
\end{aligned}$$

□

Proof of Theorem 3. From Lemmas 6 and 7, if we further assume $nh^4\to\infty$, orders in $S_{1,km}$ dominates those in $S_{2,k}$ and the leading term in cross-validation becomes Equation (17).

Define

$$A_1 = q^{-2}\sum_{k=1}^{q}\sum_{m=1}^{q} A^*_{1,km} \qquad A_2 = q^{-2}\sum_{k=1}^{q}\sum_{m=1}^{q} A^*_{2,km}$$
$$A_3 = q^{-2}\sum_{k=1}^{q}\sum_{m=1}^{q} A^*_{3,km} \qquad A_4 = q^{-2}\sum_{k=1}^{q}\sum_{m=1}^{q} A^*_{4,km}$$
$$B_1 = 2q^{-1}\sum_{k=1}^{q} B^*_{1,k} \qquad B_2 = 2q^{-1}\sum_{k=1}^{q} B^*_{2,k}.$$

Hence, Equation (12) can be written as

$$CV(h,\lambda) = A_1h^4 - A_2h^2\lambda + A_3\lambda^2 + A_4(nh^p)^{-1} + B_1\frac{h^2}{\sqrt{n}} + B_2\frac{\lambda}{\sqrt{n}} + s.o.,$$

and the leading terms of $CV(h,\lambda)$ are collected in CV_0 in Equation (17).

From Equation (17), we have

$$\begin{aligned} CV_0 &= A_1h^4 - A_2h^2\lambda + A_3\lambda^2 + A_4(nh^p)^{-1} \\ &= A_3\left(\lambda - \frac{A_2}{2A_3}h^2\right)^2 + \left(A_1 - \frac{A_2^2}{4A_3}\right)h^4 + \frac{A_4}{nh^p}. \end{aligned}$$

CV_0 is minimized when $\lambda_0 = \frac{A_2}{2A_3}h_0^2$ and $h_0^{p+4} = \frac{A_4}{4n(A_1 - A_2^2/(4A_3))}$. Hence,

$$h_0 = c_1 n^{-1/(p+4)}, \tag{25}$$
$$\lambda_0 = c_2 n^{-2/(p+4)}, \tag{26}$$

where

$$\begin{aligned} c_1 &= \left[\frac{A_4}{A_1 - A_2^2/(4A_3)}\right]^{1/(p+4)}, \\ c_2 &= \left[\frac{A_2A_4}{2A_3(A_1 - A_2^2/(4A_3))}\right]^{1/(p+4)}. \end{aligned}$$

To prove the rate of convergence of $\hat{h}$ and $\hat{\lambda}$, rewrite

$$\begin{aligned} CV(h,\lambda) &= A_3\left(\lambda - \frac{A_2h^2 - B_2n^{-1/2}}{2A_3}\right)^2 - A_3\left(\frac{A_2h^2 - B_2n^{-1/2}}{2A_3}\right)^2 \\ &\quad + A_1h^4 + A_4(nh^p)^{-1} + B_1h^2/\sqrt{n}, \\ &= A_3\left(\lambda - \frac{A_2h^2 - B_2n^{-1/2}}{2A_3}\right)^2 \\ &\quad + \left(A_1 - \frac{A_2^2}{2}\right)h^4 + (A_2B_2 + B_1)\,n^{-1/2}h^2 + \frac{A_4}{nh^p} - \frac{B_2^2}{2}n^{-1}. \end{aligned} \tag{27}$$

Minimize $CV(h,\lambda)$ w.r.t. (h,λ) in Equation (27) and we have

$$\hat{\lambda} = \frac{A_2\hat{h}^2 - B_2 n^{-1/2}}{2A_3}, \tag{28}$$

$$4(A_1 - A_2^2/2)\hat{h}^3 + 2(A_2B_2 + B_1)n^{-1/2}\hat{h} - \frac{pA_4}{n\hat{h}^{p+1}} = 0. \tag{29}$$

Let $\hat{h} = h_0 + h_1$, where h_1 is $o(h_0)$ since $\left(\hat{h} - h_0\right)/h_0 = o(1)$ and $CV(h,\lambda) = CV_0(h,\lambda) + o(CV_0)$. Substitute $\hat{h} = h_0 + h_1$, $(h_0+h_1)^{p+4} = h_0^{p+4} + (p+4)h_0^{p+3}h_1 + s.o.$ and Equation (25) into Equation (29), we obtain

$$4\left(A_1 - A_2^2/(4A_3)\right)(p+4)h_0^{p+3}h_1 + 2(A_2B_2/(2A_3) + B_1)n^{-1/2}h_0^{p+2} + s.o. = 0,$$

which gives

$$h_1 = \frac{(A_2B_2/(2A_3) + B_1)n^{-1/2}h_0^3}{2(A_1 - A_2^2/(4A_3))(p+4)h_0^4}. \tag{30}$$

Replacing h_1 with $\hat{h} - h_0$ in Equation (30) gives

$$\frac{\hat{h} - h_0}{h_0} = \frac{(A_2B_2/(2A_3) + B_1)}{2(A_1 - A_2^2/(4A_3))(p+4)c_1^2} n^{-p/(2(p+4))} = O_p(n^{-p/(2(p+4))}). \tag{31}$$

For $\hat{\lambda}$, substitute $\hat{h} = h_0 + h_1$ and $\lambda_0 = \frac{A_2}{2A_3}h_0^2$ into Equation (28) to obtain

$$\begin{aligned}
\hat{\lambda} &= A_2(h_0+h_1)^2/(2A_3) - n^{-1/2}B_2/(2A_3) \\
&= \lambda_0 + 2h_0h_1A_2/(2A_3) + h_1^2A_2/(2A_3) - h^{-1/2}B_2/(2A_3) \qquad (32)\\
&= \lambda_0 + O_p(n^{-1/2}), \qquad (33)
\end{aligned}$$

where the last line in Equation (33) follows $h_0h_1 = O(n^{-1/2})$, which can be verified by multiplying both sides of Equation (30) by h_0. □