Economics Bulletin

Volume 33, Issue 1

An Improved Local-linear Estimator For Nonparametric Regression With Autoregressive Errors

Ke Yang
University of Hartford

Abstract

In this paper we propose a modification of the local linear smoother to account for the autocorrelated errors in a nonparametric regression model with random-design. The proposed estimator has a closed-form expression and is simple to calculate. The asymptotic bias and variance of the proposed estimator are studied for AR(1) case. Compared to the standard local linear smoother, the proposed estimator retains the same design-adaptive bias but has a smaller asymptotic variance. Therefore the proposed method improves the estimation efficiency in kernel regression

I would like to thank two anonymous referees for their helpful comments and suggestions on this paper.
Citation: Ke Yang, (2013) "An Improved Local-linear Estimator For Nonparametric Regression With Autoregressive Errors", *Economics Bulletin*, Vol. 33 No. 1 pp. 19-27.
Contact: Ke Yang - kyang@hartford.edu.
Submitted: July 12, 2012. **Published:** January 08, 2013.

1 Introduction

The local polynomial regression in the presence of correlated errors has attracted much of attention in recent research. Consider the following regression model

$$Y_t = m(X_t) + \epsilon_t \quad \text{for} \quad t = 1, \cdots, T \tag{1}$$

where $m(\cdot)$ is a smooth and otherwise unknown function which is of our main interest in estimation. The error process $\{\epsilon_t, t \in \mathbb{Z}\}$ is autocorrelated and satisfies $\mathrm{E}(\epsilon_t|X_1, \cdots, X_T) = 0$ almost surely. The setting arises in many economic studies where data are gathered sequentially as times series therefore can not be assumed as independent. Marsry (1996), Masry and Fan (1997), Härdle and Tsybakov (1997) study the standard local polynomial smoother, which by construction does not take into consideration the dependence across the observations, in model 1 with random-design. Francisco-Fernndez and Vilar-Fernndez (2001) and Opsomer and Yang (2001) study the standard local polynomial regression in model (1) but with fixed-design. In both cases the local polynomial estimator has a variance proportional to the marginal variance of ϵ_t. Practitioners accustomed to correcting standard errors for dependence usually believe that adjusting for the dependence should improve the estimation accuracy, as in parametric regression. Vilar-Fernndez and Francisco-Fernndez (2002) propose a modification on local polynomial regression by including a "GLS weighting" for autocorrelation in the criterion function. The resulting estimator, however, does not improve the asymptotic properties. Xiao et al. (2003) and Martins-Filho and Yao (2009) study a two-stage "pre-whitening" procedure based on transforming the initial kernel regression residuals into a white noise process. They show that the resulting estimator reduces the variance over the conventional kernel estimator, up to the first order.[1]

In this paper we propose a natural modification of local polynomial procedure for estimating $m(\cdot)$ in model 1 that takes into consideration the correlation structure of the error terms. To illustrate the main idea, we assume that the error process $\{\epsilon_t, t \in \mathbb{Z}\}$ follows an AR(1) type correlation structure. The proposed procedure is highly intuitive in its construction and is easy to compute, compared to the "pre-whitening" method. More importantly, the proposed kernel estimator has a smaller variance compared to the conventional local polynomial estimator while maintain the same bias, up to the first order. In this sense it is asymptotically more efficient. The paper is structured as follows. The proposed procedure is introduced in Section 2 and its asymptotic properties is studied in Section 3. We conclude the paper in Section 4.

2 The Model and the Estimator

Suppose in model (1) the error process $\{\epsilon_t, t \in \mathbb{Z}\}$ follows $\epsilon_t = \rho\epsilon_{t-1} + e_t$ where $|\rho| < 1$ and $\{e_t, t \in \mathbb{Z}\}$ is a white-noise process with finite variance σ_e^2. Let $\underline{X} = (X_1, \cdots, X_T)'$

[1] There is a related literature regarding kernel regression on longitudinal or panel data. See Severini and Staniswalis (1994), Ruckstuhl et al. (2000), Lin and Carroll (2000), Wang (2003) and Chen and Jin (2005), among others.

and similarly for $\underline{Y}$, $\underline{\epsilon}$. In addition let $\underline{m}(\underline{X}) = (m(X_1), \cdots, m(X_T))'$. The covariance matrix of the error process is $E(\underline{\epsilon}\underline{\epsilon}') = \sigma_e^2\Sigma$ where $\Sigma = [\nu_{jk}]_{j,k=1}^T$ with $\nu_{jk} = \frac{1}{1-\rho^2}\rho^{|j-k|}$. Since Σ is positive definite, there exists a nonsingular matrix $\underline{P}$ such that $\underline{P}'\underline{P} = \underline{\Sigma}^{-1}$. The standard local polynomial regression aims at the p^{th} polynomial expansion of the regression function at a local point, $m(x) \approx G_p(X-x)'\beta$, where $G_p(v) = \{1, v, \cdots, v^p\}'$ and $\beta = \{\beta_0, \cdots, \beta_p\}'$ is a vector of Taylor expansion coefficients. Let h denote the bandwidth and $K(\cdot)$ a symmetric kernel function defined on a compact support, normalized without loss of generality to have unit variance. Define $K_h(v) = h^{-1}K(v/h)$ and $\underline{G}_p(x) = \{G_p(X_1 - x), G_p(X_2 - x)), \cdots, G_p(X_T - x)\}'$. To focus on the main idea and keep the paper concise, we let X_t be a scalar. Meanwhile we also concentrate on the local linear regression setting, i.e. $p = 1$, so the subscript for the order of polynomial in $\underline{G}_p(\cdot)$ can be skipped. In the standard local linear regression, $m(x)$ is estimated by $\tilde{m}(x) = \tilde{\beta}_0(x)$ with $\tilde{\beta}(x) = \{\tilde{\beta}_0(x), \tilde{\beta}_1(x)\}$ defined as

$$\tilde{\beta}(x) = \{\underline{G}(x)'\underline{K}(x)\underline{G}(x)\}^{-1}\underline{G}(x)'\underline{K}(x)\underline{Y}$$

where $\underline{K}(x) = diag\{K_h(X_t - x)\}_{t=1}^T$. Martins-Filho and Yao (2009) show that the $\tilde{m}$ has a variance

$$Var_{\underline{X}}(\tilde{m}(x)) = \frac{1}{Th}\frac{\sigma_\epsilon^2\gamma(0)}{f(x)} + o_p(\frac{1}{Th}) = \frac{1}{Th}\frac{\sigma_e^2\gamma(0)}{(1-\rho^2)f(x)} + o_p(\frac{1}{Th}) \tag{2}$$

Note that when the errors are highly correlated, i.e. ρ is close to 1, the variance given above can be very large. In order to utilize the dependence structure in the model, we propose to estimate $m(x)$ by $\hat{m}(x) = \hat{\beta}_0$ with $\hat{\beta}(x) = \{\hat{\beta}_0(x), \hat{\beta}_1(x)\}$ defined as

$$\hat{\beta}(x) = \left\{\underline{G}(x)'\underline{\mathbf{1}}_x\underline{P}'\underline{K}(x)\underline{P}\underline{\mathbf{1}}_x\underline{G}(x)\right\}^{-1}\underline{G}(x)'\underline{\mathbf{1}}_x\underline{P}'\underline{K}(x)\underline{P}\underline{\mathbf{1}}_x\underline{Y} = (B_T)^{-1}C_T \tag{3}$$

where the $\underline{\mathbf{1}}(x) = diag\{\mathbf{1}(\frac{X_t - x}{h})\}_{t=1}^T$, and $\mathbf{1}(\cdot)$ is an indicator function defined on the same support as the kernel function, $K(\cdot)$. Note that in (3), the indicator matrix $\underline{\mathbf{1}}_x$ and kernel matrix $\underline{K}(x)$ are placed on either side of the correlation matrix $\underline{P}$ in order to select only those elements in $\underline{P}$ corresponding to observations within the same bandwidth with the local point x on which $m(\cdot)$ is estimated. Meanwhile, note that in most practical situations $\hat{\beta}(\cdot)$ is not feasible because the matrix $\underline{P}$ is unknown and needs to be estimated. A method of moment estimator for $\underline{P}$ can be obtained by substituting ρ by its method of moment estimator $\hat{\rho}$, i.e.

$$\hat{\rho} = \frac{\sum_{t=1}^{T-1}\hat{\epsilon}_t\hat{\epsilon}_{t+1}}{\sum_{t=1}^{T}\hat{\epsilon}_t^2} \tag{4}$$

where $\hat{\epsilon}_t = Y_t - \tilde{m}(x_t)$, $1 \leq t \leq T$, are residuals from a standard local linear regression. A feasible version of the proposed estimator for $m(x)$ is $\hat{m}^*(x) = \hat{\beta}_0^*(x)$ with $\hat{\beta}^*(x) = \{\hat{\beta}_0^*(x), \hat{\beta}_1^*(x)\}$ defined as

$$\hat{\beta}^*(x) = \left\{\underline{G}(x)'\underline{\mathbf{1}}_x\underline{\hat{P}}'\underline{K}(x)\underline{\hat{P}}\underline{\mathbf{1}}_x\underline{G}(x)\right\}^{-1}\underline{G}(x)'\underline{\mathbf{1}}_x\underline{\hat{P}}'\underline{K}(x)\underline{\hat{P}}\underline{\mathbf{1}}_x\underline{Y} = (\hat{B}_T)^{-1}\hat{C}_T \tag{5}$$

where it is assumed that $(\hat{B}_T)^{-1}$ exists.

3 Asymptotic Properties

The asymptotic analysis of the proposed estimator is focused on the bias and variance approximation. We assume that the error process $\{\epsilon_t\}$ is independent of the process $\{X_t\}$. To facilitate the asymptotic analysis, the following assumptions are made on the kernel function, the bandwidth h, the independent variable X_t and the conditional mean function $m(x)$.

Assumption 1 *$KernelK(\cdot)$ is a symmetric and bounded kernel density function defined on $[-1/2, 1/2]$ and it has a unit variance. Let $\mu(r) = \int \phi^r K(\phi)d\phi$, $\mu_{\mathbf{1}}(r) = \int \phi^r \mathbf{1}(\phi)d\phi$ and $\gamma(r) = \int \phi^r K^2(\phi)d\phi < \infty$, $r = 1, 2, \cdots$, then we have $\mu(r) = \mu_{\mathbf{1}}(r) = 0$ for odd r and $\mu(2) = 1$;*

Assumption 2 *As $T \to \infty$ we have $Th \to \infty$ and $h \to 0$;*

Assumption 3 *The second derivative of $m(\cdot)$ exists and it is continuous, denoted as $m^{''}(\cdot)$;*

Assumption 4 *The regressor $\{X_t, t \in \mathbb{Z}\}$, is a stationary process. The marginal density of X_t, denoted by $f(x)$ and the joint density, denoted by $f_{t,t-1}(X_{t,}X_{t-1})$ are continuous and bounded both from above and away from zero.*

The asymptotic bias and variance of the infeasible version of the estimator are given in Theorem 3.1.

Theorem 3.1 *Given assumptions $\boldsymbol{A}1$ to $\boldsymbol{A}4$ and suppose x is an interior point on the support of $f(\cdot)$, the bias and variance of the proposed estimator $\hat{m}(x)$ can be approximated as follows:*

The bias is:

$$E_{\underline{X}}\left\{\hat{m}(x) - m(x)\right\} = \frac{1}{2}h^2\mu(2)m'' + o_p(h^2) \tag{6}$$

The variance is:

$$Var_{\underline{X}}\left(\hat{m}(x)\right) = (Th)^{-1}\frac{\sigma_e^2\gamma(0)}{f(x)} + o_p(\frac{1}{Th}) \tag{7}$$

The proof is given in the appendix. We make three remarks on this result.

Remark 1: The leading term in the bias of $\hat{m}(x)$ in (6) is the same as that of the standard local linear estimator. Meanwhile the variance of $\hat{m}(x)$ in (7) is proportional to σ_e^2 , while the conventional local linear smoother has a variance proportional to $\sigma_\epsilon^2 = \sigma_e^2/(1-\rho^2)$, which is strictly greater than σ_e^2 except when $\rho = 0$, i.e. errors are independent. This argument holds for all the values of x, regardless of the values of $m''(\cdot)$ and ρ. Hence $\hat{m}(\cdot)$ is more efficient in the sense of achieving smaller mean square errors (MSE) uniformly. The relative efficiency gain measured by MSE increases with correlation level.

Remark 2: The asymptotic variance of the "GLS-weighting" estimator proposed in Vilar-Fernndez and Francisco-Fernndez (2002) is proportional to $\frac{\sigma_e^2}{(1-\rho)^2}$, which is same as that of the standard local linear estimator and strictly larger than variance of the proposed estimator. In this sense our proposed estimator is more efficient asymptotically.

Remark 3: The "pre-whitening" procedure in Xiao et al. (2003) and Martins-Filho and Yao (2009) has the same asymptotic bias and variance as $\hat{m}(\cdot)$ when an under-smoothing bandwidth, h_0, is used in the first-stage regression, i.e. $h_0/h \to 0$ as $T \to \infty$. However, when $h_0 = h$, the asymptotic bias of pre-whitening estimator is much more complex and it depends on the values of $m''(x)$ and ρ. Numerically it can potentially be greater than the bias in (6). In this sense, the "pre-whining" estimator does not outperform the conventional standard local linear estimator uniformly.

To derive the mean square properties of the feasible estimator, we need $\hat{\rho}$ to converge in probability to ρ, which is given in the following proposition.

Propsition 3.2 *Under assumption A1 to A4, we have*

$$\hat{\rho} \longrightarrow \rho \quad as \quad T \longrightarrow \infty \quad with\ probability \quad 1.$$

Proof The argument to establish the convergence of $\hat{\rho}$ follows Altman (1990, Theorem 3). ∎

In order to derive the asymptotic properties of $\hat{m}^*(x)$, we need the following assumption on the bandwidth h.

Assumption 5 $Th^5 \longrightarrow C < \infty$ *as* $n \longrightarrow \infty$;

Using the Theorem 3.1 and Proposition 3.2 we can derive the bias and variance for $\hat{m}^*(x)$.

Theorem 3.3 *Given assumptions A1 to A5 and suppose x is a interior point on the support of $f(\cdot)$, we have*

$$\sqrt{Th}\,(\hat{m}^*(x) - \hat{m}(x)) \longrightarrow 0 \quad as \quad T \longrightarrow \infty, \quad with\ probability\ 1.$$

The proof is given in the appendix.

4 Conclusion

In this paper we propose a modification of the local polynomial smoother to accommodate the dependence structure in a autoregressive error process. One advantage of the proposed estimator is its simplicity. It has a closed form expression and analogous to the classical weighted least square type estimator in parametric regression on autoregressive data. In the case that the error process is AR(1), the asymptotic analysis shows that the proposed estimator improves the estimation efficiency over the standard local polynomial smoother and the improvement can be large depending on the autocorrelation function. Extending the proposed method to a local polynomial smoothing of order p , or to a regression model with more generally specified error process, like ARMA(p,q), is conceptually straightforward but further efforts are needed to establish the asymptotic properties of the proposed estimator.

References

Altman, N.S. (1990), "Kernel smoothing of data with correlated errors." *Journal of the American Statistical Association*, 85, 749–759.

Chen, K. and Z. Jin (2005), "Local polynomial regression analysis of clustered data." *Biometrika*, 92 (1), 59 – 74.

Francisco-Fernndez, Maro and Juan Manuel Vilar-Fernndez (2001), "Local polynomial regression estimation with correlated errors." *Communications in statistics, theory and methods*, 30 (7), 1271–1293.

Härdle, W. and A. Tsybakov (1997), "Local polynomial estimators of the volatility function in nonparametric autoregression." *Journal of Econometrics*, 81, 223–242.

Lin, X. and R.J. Carroll (2000), "Nonparametric function estimation for clustered data when the predictor is measured without/with error." *Journal of the American Statistical Association*, 95, 520–534.

Marsry, Elias (1996), "Multivariateregression estimation localpolynomialfitting for time-series." *Stochastic Processes and their Applications*, 65, 81–101.

Martins-Filho, C. and F. Yao (2009), "Nonparametric regression estimation with general parametric error covariance." *Journal of Multivariate Analysis*, 100, 309–333.

Masry, E. and J. Fan (1997), "Local polynomial estimation of regression functions for mixing processes." *Scandinavian Journal of Statistics*, 24, 165–179.

Opsomer, Wang Y., J.D. and Y. Yang (2001), "Nonparametric regression with correlated errors." *Statistical Science*, 16, 134–153.

Ruckstuhl, A.F., A.H. Welsh, and R.J. Carroll (2000), "Nonparametric function estimation of the relationship between two repeatedly measured variables." *Statistica Sinica*, 10, 51–71.

Severini, T.A. and J.G. Staniswalis (1994), "Quasi-likelihood estimation in semiparametric models." *Journal of the American Statistical Association*, 89, 501–511.

Vilar-Fernndez, J.M. and M Francisco-Fernndez (2002), "Local polynomial regression smoothers with ar-error structure." *Test*, 11, 439–464.

Wang, N. (2003), "Marginal nonparametric kernel regression accounting for within-subject correlation." *Biometrika*, 90, 43–52.

Xiao, Z., O.B. Linton, R.J. Carroll, and E. Mammen (2003), "More efficient local polynomial estimation in nonparametric regression with autocorrelated errors." *Journal of the American Statistical Association*, 98, 980–992.

A Technical Appendix

A.1 proof of theorem 1

Proof First note that $\hat{\beta}$is the solution to the following equation

$$L_T(\hat{\beta}) = \frac{1}{T}\underline{G}(x)'\underline{\mathbf{1}}_x\underline{P}'\underline{K}_x\underline{P\mathbf{1}}_x(\underline{Y} - \underline{G}(x)\hat{\beta}) = 0$$

and the proposed estimator satisfies $\hat{\beta} - \beta = \frac{1}{T}(B_T)^{-1}L_T(\beta)$. In the following we define $\mathbf{1}_t = \mathbf{1}(X_t - x)$, $K_t = K_h(\frac{X_t - x}{h})$ and $z_t = X_t - x$. Direct calculation gives the elements in $\frac{1}{T}B_T$ as follows:

$$\begin{aligned}
\frac{1}{T}(B_T)_{1,1} &= \frac{1}{T}(1-\rho^2)K_1\mathbf{1}_1^2 + \frac{1}{T}\sum_{t=2}^{T} K_t(\mathbf{1}_t - \rho\mathbf{1}_{t-1})^2 \\
&\simeq \int \frac{1}{h}K(\frac{X_t - x}{h})\left\{\mathbf{1}(\frac{X_t - x}{h}) - \rho\mathbf{1}(\frac{X_{t-1} - x}{h})\right\}^2 f(X_t)f(X_{t-1})dX_t dX_{t-1} \\
&= \int K_t\mathbf{1}_t f(X_t)dX_t - 2\rho\int K_t\mathbf{1}_t\mathbf{1}_{t-1}f(X_t)f(X_{t-1})dX_t dX_{t-1} + \\
&\qquad + \rho^2\int K_t\mathbf{1}_{t-1}f(X_t)f(X_{t-1})dX_t dX_{t-1} \\
&= f(x) - 2h\rho f^2(x) + h\rho^2 f^2(x) + O(h)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{T}(B_T)_{1,2} = (B_T)_{2,1} &= \frac{1}{T}(1-\rho^2)z_1K_1\mathbf{1}_1^2 - \frac{1}{T}\sum_{t=2}^{T} K_t(\mathbf{1}_t - \rho\mathbf{1}_{t-1})(-\rho z_{t-1}\mathbf{1}_{t-1} + z_t\mathbf{1}_t) \\
&= o(h)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{T}(B_T)_{2,2} &= \frac{1}{T}(1-\rho^2)z_1^2K_1\mathbf{1}_1^2 + \frac{1}{T}\sum_{t=2}^{T}(-\rho z_{t-1}\mathbf{1}_{t-1} + z_t\mathbf{1}_t)^2K_t \\
&= \int z_t^2\mathbf{1}_t^2K_tf(X_t)dX_t - 2\rho\int z_t\mathbf{1}_tK_tz_{t-1}\mathbf{1}_{t-1}f(X_t)f(X_{t-1})dX_tdX_{t-1} + \\
&\qquad + \rho^2\int K_tz_{t-1}f(X_t)f(X_{t-1})dX_tdX_{t-1} \\
&= h^2\mu(2)f(x) + o(h^2)
\end{aligned}$$

Hence,

$$\frac{1}{T}\underline{B}_T = \begin{bmatrix} f(x) - 2h\rho f^2(x) + h\rho^2 f^2(x), & o(h) \\ o(h), & h^2\mu(2)f(x) + o(h) \end{bmatrix}$$

Then, we look at $E_{\underline{X}}\{(L_T(\beta)\} = \frac{1}{T}\underline{G}(x)^{'}\underline{\mathbf{1}_x P^T} \underline{K_x} \underline{P\mathbf{1}_x}(\underline{Y} - \underline{G}(x)^{'}\underline{\beta})$. Replacing each $m(X_t)$ with its second-order Taylor expansion around point x, we have

$$\underline{m}(\underline{X}) = \underline{G}(x)^{'}\beta + (\underline{X} - x)^2[\frac{1}{2}m''(x)] + O((\underline{X} - x)^3).$$

Then the term $E\{(L_T(\beta)|\underline{X}\}$can be written as

$$= \frac{1}{T}\underline{G}(x)^{'}\mathbf{1}_x\underline{P^{'}KP}\mathbf{1}_x\left\{(\underline{X} - x)^2[\frac{1}{2}m''(x)] + O\{(\underline{X} - x)^3\}\right\}$$
$$\simeq [\frac{1}{2}m''(x)]\frac{1}{T}\left[\begin{array}{c} \sum_{t=2}^{T} K_t[-\rho\mathbf{1}_t Z_{t-1}^2\mathbf{1}_{t-1} + \rho^2 Z_{t01}^2\mathbf{1}_{t-1}^2 + Z_t^2\mathbf{1}_t^2 + \rho Z_t^2\mathbf{1}_t\mathbf{1}_{t-1}] \\ \sum_{t=2}^{T} K_t[-\rho Z_t^2\mathbf{1}_t Z_{t-1}\mathbf{1}_{t-1} + \rho^2 Z_{t-1}^3\mathbf{1}_{t-1}^3 + Z_t^3\mathbf{1}_t^3 + \rho Z_t^2\mathbf{1}_t\mathbf{1}_{t-1}] \end{array}\right] + O(h^3)$$
$$\simeq [\frac{1}{2}m''(x)]\left[\begin{array}{c} \mu(2)f(x) + O(h) \\ h\rho\mu(2)f(x) + o(h) \end{array}\right] + O(h^3)$$

Direct calculation gives that the bias of $\hat{m}(x)$ is[2]

$$E_{\underline{X}}(\hat{m}(x) - m(x)) \simeq (1,0)(\frac{1}{T}B_T)^{-1}E_X(L_T(\beta)) = \frac{1}{2}h^2\mu(2)m''(x) + o_p(h^2)$$

Now we look at the variance of $\hat{m}(x)$ conditional on $\underline{X}$. First note that

$$Cov_{\underline{X}}\left\{\frac{1}{T}\underline{G}(x)^{'}\underline{\mathbf{1}}_x\underline{P}^{'}\underline{K}_x\underline{P\mathbf{1}}_x\underline{Y}\right\} = \sigma_e^2\frac{1}{T}\underline{G}(x)^{'}\underline{\mathbf{1}}_x\underline{P}^{'}\underline{K}_x\underline{P\mathbf{1}}_x\Sigma\underline{\mathbf{1}}_x\underline{P}^{'}\underline{K}_x\underline{P\mathbf{1}}_x\underline{G}(x)\frac{1}{T}$$

Since $\underline{P}^{'}\underline{P} = \Sigma^{-1}$, so the middle part of the above matrix $\underline{P\mathbf{1}}_x\Sigma\underline{\mathbf{1}}_x\underline{P}^{'}$ becomes a identity matrix. Therefore, we have

$$Cov_{\underline{X}}(\cdot) = \sigma_e^2\frac{1}{T}\underline{G}(x)^{'}\underline{\mathbf{1}}_x\underline{P}^{'}\underline{K}_x\underline{K}_x\underline{P\mathbf{1}}_x\underline{G}(x)\frac{1}{T} = \frac{\sigma_e^2}{T^2}AA^{'}$$

Direct calculation shows that the $(i,j)^{th}$, $i,j = 1,2$, element of the matrix $Cov_X(\cdot)$, denoted by $C_{i,j}$, are as follows:

$$C_{1,1} = \frac{\sigma_e^2}{T^2}(1-\rho^2)K_1^2\mathbf{1}_1^2 + \frac{\sigma_e^2}{T^2}\sum_{t=2}^{T}K_t^2(\mathbf{1}_t - \rho\mathbf{1}_{t-1}) \simeq \frac{1}{Th}\sigma_e^2\gamma(0)f(x)$$
$$C_{1,2} = C_{2,1} = \frac{\sigma_e^2}{T^2}(1-\rho^2)z_1K_1^2\mathbf{1}_1^2 + \frac{\sigma_e^2}{T^2}\sum_{t=2}^{T}K_t^2\left(\mathbf{1}_t - \rho\mathbf{1}_{t-1}\right)\left(-\rho z_{t-1}\mathbf{1}_{t-1} + z_t\mathbf{1}_t\right)$$
$$\simeq \frac{\sigma_e^2}{Th}[\gamma(1)f(x) + o(h) + O(h)]$$
$$C_{2,2} = \frac{\sigma_e^2}{T^2}(1-\rho^2)z_1^2K_1^2\mathbf{1}_1^2 + \frac{\sigma_e^2}{T^2}\sum_{t=2}^{T}\left(-\rho z_{t-1}\mathbf{1}_{t-1} + z_t\mathbf{1}_t\right)^2K_t^2$$
$$= \frac{\sigma_e^2}{Th}[\gamma(2)f(x) + o(h)]$$

[2] The determinant of B_T is $det(B_T) = (B_T)_{1,1}(B_T)_{2,2} - (B_T)_{2,1}(B_T)_{1,2}$.

So in matrix form, the covariance of the above term is

$$Cov_{\underline{X}}(\cdot) = \frac{\sigma_e^2}{Th} f(x) \begin{bmatrix} \gamma(0), & \gamma(1) \\ \gamma(1), & \gamma(2) \end{bmatrix} + o_p(\frac{1}{Th})$$

Now, the variance of $\hat{\beta}_0$ is

$$\begin{aligned} Var_{\underline{X}}(\hat{\beta_0}) &= \{1,0\}\{B_T\}^{-1} Cov_X(\frac{1}{T} G_{p,h}(\underline{X} - x\underline{e}_T)\underline{I}_{x,h}\underline{P}^{'}\underline{KPI}_{x,h}\underline{Y})\{B_T\}^{-1}\{1,0\}^{'} \\ &= \frac{\sigma_e^2}{Th}\frac{v(0)}{f(x)} + o_p(1/Th) \end{aligned}$$

∎

A.2 Proof of Theorem 3.3

Proof First we let $W^{-1} = \underline{I}_{x,h}\underline{P}^{'}\underline{KPI}_{x,h}$ and $\hat{W}^{-1} = \underline{I}_{x,h}\underline{\hat{P}}^{'}\underline{K\hat{P}I}_{x,h}$; Define $\underline{G}^{+1} = \{(X_1 - x)^2, \cdots, (X_T - x)^2\}^{'}$. From the definition of $\hat{\beta}$ and $\hat{\beta}^*$, we have that

$$\begin{aligned} \hat{\beta} - \hat{\beta}^* &= (\underline{G}^{'}W^{-1}\underline{G})^{-1}\underline{G}^{'}W^{-1}Y - (\underline{G}^{'}\hat{W}^{-1}\underline{G})^{-1}\underline{G}^{'}\hat{W}^{-1}Y \\ &= (\underline{G}^{'}W^{-1}\underline{G})^{-1}\underline{G}^{'}W^{-1}(\underline{G}^{+1}\beta_2 + o(h^3) + \underline{\epsilon}) - (\underline{G}^{'}\hat{W}^{-1}\underline{G})^{-1}\underline{G}^{'}\hat{W}^{-1}(\underline{G}^{+1}\beta_2 + o(h^3) + \underline{\epsilon}) \\ &= \beta_2\left[(\underline{G}^{'}W^{-1}\underline{G})^{-1}\underline{G}^{'}W^{-1} - (\underline{G}^{'}\hat{W}^{-1}\underline{G})^{-1}\underline{G}^{'}\hat{W}^{-1}\right]\underline{G}^{+1} \\ &\qquad + \left[(\underline{G}^{'}W^{-1}\underline{G})^{-1}\underline{G}^{'}W^{-1} - (\underline{G}^{'}\hat{W}^{-1}\underline{G})^{-1}\underline{G}^{'}\hat{W}^{-1}\right]\underline{\epsilon} + o_p(h^3) \\ &= \Delta_1 + \Delta_2 \end{aligned}$$

From proof of theorem 3.1, we get

$$\begin{aligned} E_{\underline{X}}(\hat{\beta} - \hat{\beta}^*) &= \Delta_1 = h^2 o_p(1) \\ var_{\underline{X}}(\hat{\beta} - \hat{\beta}^*) &= \Delta_2\Delta_2^{'} = \frac{1}{Th} \cdot o_p(1) \end{aligned}$$

Combine this with assumption 5, we have

$$\begin{aligned} E_{\underline{X}}(\sqrt{Th}(\hat{\beta} - \hat{\beta}^*)) &= o_p(1) \\ var_{\underline{X}}(\sqrt{Th}(\hat{\beta} - \hat{\beta}^*)) &= o_p(1) \end{aligned}$$

Therefore, we have $\sqrt{Th}(\hat{\beta} - \hat{\beta}^*) = o_p(1)$.

∎