Covariate Adjusted Correlation Analysis via Varying Coefficient Models

DAMLA ŞENTÜRK

Department of Statistics, Pennsylvania State University

HANS-GEORG MÜLLER

Department of Statistics, University of California, Davis

ABSTRACT. We propose covariate adjusted correlation (Cadcor) analysis to target the correlation between two hidden variables that are observed after being multiplied by an unknown function of a common observable confounding variable. The distorting effects of this confounding may alter the correlation relation between the hidden variables. Covariate adjusted correlation analysis enables consistent estimation of this correlation, by targeting the definition of correlation through the slopes of the regressions of the hidden variables on each other and by establishing a connection to varying-coefficient regression. The asymptotic distribution of the resulting adjusted correlation estimate is established. These distribution results, when combined with proposed consistent estimates of the asymptotic variance, lead to the construction of approximate confidence intervals and inference for adjusted correlations. We illustrate our approach through an application to the Boston house price data. Finite sample properties of the proposed procedures are investigated through a simulation study.

Key words: additive distortion model, asymptotic normality, confidence interval, confounding, linear regression, martingale CLT, multiplicative distortion model, smoothing, varying coefficient model.

Running Title: Covariate Adjusted Correlation

1. Introduction

We address the problem of estimating the correlation between two variables which are distorted by the effect of a confounding variable. For identifiability, we assume that the mean distorting effect of the confounder corresponds to no distortion. We consider the multiplicative effects model where the actual variables (Y, X) are observed after being multiplied by a smooth unknown function of the confounder, leading to the observations

$$\tilde{Y}_i = \psi(U_i)Y_i, \qquad \text{and} \qquad \tilde{X}_i = \phi(U_i)X_i. \tag{1}$$

Here (Y_i, X_i) represent the unobserved realizations of the actual variables, $\psi(\cdot)$, $\phi(\cdot)$ are unknown smooth functions of the confounder with observed values U_i, and $(\tilde{Y}_i, \tilde{X}_i)$ are the distorted observations of the original variables for a sample size of n. The nature of the relationship of the confounding variable U with the underlying variables will often be unknown, implying that $\psi(\cdot)$ and $\phi(\cdot)$ in (1) are unknown functions.

The identifiability condition of no average distortion can be expressed as

$$E\{\psi(U)\} = 1, \qquad E\{\phi(U)\} = 1. \tag{2}$$

We also assume that (Y_i), (X_i), (U_i) are independent and identically distributed for $i = 1, \ldots, n$, where U is independent of Y and X.

One example is the Boston house price data of Harrison & Rubinfeld (1978), where finding the correlation relation between crime rate and house prices is of interest. However, a confounder affecting both variables is the proportion of population of lower educational status. For such data, model (1) provides a general and sensible way to describe this confounding as we shall demonstrate.

The simultaneous dependence of the original variables on the same confounder may lead to artificial correlation relations which do not exist between the actual hidden variables whose correlation we want to infer. To illustrate how drastically the multiplicative distorting effects of the confounder may change the correlation between the underlying variables even under the identifiability condition of no average distortion, consider the following example. The underlying variables Y vs. X, simulated from a bivariate normal distribution with $\rho_{(Y,X)} = 0.5$, and the distorted versions

$\tilde{Y}$ vs. $\tilde{X}$, where the distortion is multiplicative as given in (1) through smooth, almost linear functions of a uniformly distributed confounder, have been plotted in Figure 1. Detailed explanations are given in the simulation study in Section 5.2. For this data, the sample correlation between the underlying variables is $\hat{\rho}_{(Y,X)} = 0.4924$, whereas for the distorted variables the sample correlation is negative, $\hat{\rho}_{(\tilde{Y},\tilde{X})} = -0.4552$.

FIGURE 1 ABOUT HERE

A central goal of this paper is consistent estimation and inference for $\rho_{(Y,X)}$, the correlation coefficient of the hidden variables (Y, X), given the observations of the confounding variable U_i and the distorted observations $(\tilde{Y}_i, \tilde{X}_i)$ in (1). We refer to model (1), (2) as multiplicative distortion model. Adjustment for confounding variables per se is a classical problem. We start by investigating a sequence of nested models, for all of which standard adjustment methods to obtain consistent estimates of the correlation already exist.

First, consider an additive distortion model, $\tilde{Y} = Y + \psi_a(U)$ and $\tilde{X} = X + \phi_a(U)$, with identifiability constraints $E\{\psi_a(U)\} = E\{\phi_a(U)\} = 0$, for the distorting effects of U to average out to 0. A simple adjustment method for the consistent estimation of $\rho_{(Y,X)}$ in the additive distortion model is to use an estimate of $\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$, where $\tilde{e}_{W_1W_2}$ is the set of errors from the nonparametric regression of W_1 on W_2 (referred to as nonparametric partial correlation). However, as we show in Appendix D, under (1), (2), the estimate of $\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$ is targeting the value

$$\xi_1 = \rho_{(Y,X)}\Delta, \tag{3}$$

where $\Delta = E\{\psi(U)\phi(U)\}/[\sqrt{E\{\psi^2(U)\}}\sqrt{E\{\phi^2(U)\}}]$. Noting that Δ can assume any real value in the interval $(0, 1]$, this simple adjustment, while working for the special case of an additive distortion model, fails for the multiplicative distortion model.

The second model considered is a special case of the additive distortion model, where the distorting functions $\psi_a(\cdot)$ and $\phi_a(\cdot)$ are linear functions of U. In this case, a consistent estimate of $\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$ will also be consistent for $\rho_{(Y,X)}$ where $e_{W_1W_2}$ is the set of errors from the least squares regression of W_1 on W_2. This simple adjustment method is also known as the partial correlation of $\tilde{Y}$ and $\tilde{X}$, adjusted for U (Pearson, 1896). This popular adjustment however fails for the multiplicative distortion model, since under (1), (2), the

target value ξ_2 of the estimate of $\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$ will generally not satisfy $\xi_2 = \rho_{(Y,X)}$. Indeed it holds that $\xi_2 = \xi_1$, where ξ_1 is as given in (3) (see Appendix E). This adjustment method therefore is fraught with the same bias problem as the nonparametric partial correlation adjustment discussed above.

What if we ignore the distorting effects of the confounder U on $(\tilde{Y}, \tilde{X})$? In this case we would simply use the regular correlation $\rho_{(\tilde{Y},\tilde{X})}$ in order to target $\rho_{(Y,X)}$. As we will show in Appendix F, under (1), (2), this correlation estimate is targeting the value

$$\xi_3 = \frac{E\{\phi(U)\psi(U)\}E(XY) - E(X)E(Y)}{\sqrt{\text{var}\{\phi(U)\}E(X^2) + \text{var}(X)}\sqrt{\text{var}\{\psi(U)\}E(Y^2) + \text{var}(Y)}}. \tag{4}$$

Now if $\psi(\cdot) \equiv \phi(\cdot) \equiv 1$, i.e., there is no confounding, then we immediately see that $\rho_{(\tilde{Y},\tilde{X})} = \rho_{(Y,X)}$, so that in this case of no confounding this estimate is on target. However, if $\psi(\cdot)$ and $\phi(\cdot)$ do not equal one, then we find that ξ_3 can assume any real value in the interval $[-1, 1]$. Therefore, arbitrarily large biases can result if one estimates a correlation while ignoring the confounding.

Another straightforward approach would be to apply logarithmic transformations to $\tilde{Y}$ and $\tilde{X}$, so as to change the effect of the distorting functions $\psi(\cdot)$ and $\phi(\cdot)$ from multiplicative to additive. We would then use the nonparametric partial correlation adjustment method to estimate $\rho_{\{\log(\tilde{Y}),\log(\tilde{X})\}}$ consistently. However, this ad hoc solution might fall short since the observed $\tilde{Y}$ and $\tilde{X}$ might not necessarily be positive. Furthermore, one may be interested in the correlation between the untransformed variables which is not easy to recover from that of the transformed variables.

Since the available adjustment methods fail to adjust properly for the distorting effects of the confounder U in the multiplicative distortion model, a new adjustment method for correlations needs to be developed, a problem that we address in this paper. Our starting point is the equality

$$\rho_{(Y,X)} = \text{sign}(\gamma_1)\sqrt{\gamma_1\eta_1} = \text{sign}(\eta_1)\sqrt{\gamma_1\eta_1}, \tag{5}$$

where γ_1 and η_1 are the slopes from the linear regressions of Y on X and X on Y respectively, and $\rho_{(Y,X)}$ is the underlying targeted correlation. We then propose an estimate of $\rho_{(Y,X)}$, based on pilot estimates of γ_1 and η_1. Assuming that the linear models given in (6)

below hold between the underlying variables Y and X, the proposed general estimation method provides a consistent estimate for $\rho_{(Y,X)}$ not only under a multiplicative distortion model, but under all three distortion models outlined above (Şentürk & Müller, 2003a). This is one of the major attractions of the proposed adjustment since in most applications the specific nature of the distortion will be unknown. The asymptotic distribution of the resulting covariate adjusted correlation (Cadcor) estimates is established. This main result, combined with proposed consistent estimates for the asymptotic variance, is then applied for the construction of approximate confidence intervals for the correlation coefficient.

The paper is organized as follows. In Section 2 we describe the model in more detail and explore the relationship to varying coefficient models. In Section 3 we introduce the covariate adjusted correlation (Cadcor) estimates and present results on asymptotic inference. Consistent estimates for the asymptotic variance, as needed for the implementation of inference procedures, are derived in Section 4. Applications of the proposed method to the Boston housing data are discussed in Section 5, where we also present some simulation results. The proofs of the main results are assembled in Section 6, followed by an Appendix with technical conditions and auxiliary results.

2. Covariate adjustment via varying coefficient regression

We assume the following regression relations between the unobserved variables Y and X:

$$\begin{aligned} Y_i &= \gamma_0 + \gamma_1 X_i + e_{YX,i}, \\ X_i &= \eta_0 + \eta_1 Y_i + e_{XY,i}, \end{aligned} \tag{6}$$

where $e_{YX,i}$ is the error term such that $E(e_{YX,i}) = 0$ with constant variance σ^2_{YX}, and $e_{XY,i}$ with $E(e_{XY,i}) = 0$ and constant variance σ^2_{XY}. Our goal is to use estimates of γ_1 and η_1 (Şentürk & Müller, 2003a) to arrive at an estimate for $\rho_{(Y,X)}$ via (5).

The regression for observed variables leads to

$$\begin{aligned} E(\tilde{Y}_i|\tilde{X}_i, U_i) &= E\{Y_i\psi(U_i)|\phi(U_i)X_i, U_i\} \\ &= \psi(U_i)E\{\gamma_0 + \gamma_1 X_i + e_{YX,i}|\phi(U_i)X_i, U_i\}. \end{aligned}$$

Assuming that $E(e_{YX,i}) = 0$ and that (e_{YX}, U, X) are mutually independent, the model reduces to

$$\begin{aligned} E(\tilde{Y}_i|\tilde{X}_i, U_i) &= \psi(U_i)\gamma_0 + \psi(U_i)\gamma_1 \frac{\phi(U_i)X_i}{\phi(U_i)} \\ &= \beta_0(U_i) + \beta_1(U_i)\tilde{X}_i, \end{aligned}$$

defining the functions

$$\beta_0(u) = \gamma_0\psi(u), \qquad \beta_1(u) = \gamma_1\frac{\psi(u)}{\phi(u)}. \tag{7}$$

This leads to

$$\tilde{Y}_i = \beta_0(U_i) + \beta_1(U_i)\tilde{X}_i + \psi(U_i)e_{YX,i},$$

corresponding to a multiple varying coefficient model, i.e. an extension of regression and generalized regression models where the coefficients are allowed to vary as a smooth function of a third variable (Hastie & Tibshirani, 1993). For varying coefficient models, Hoover *et al.* (1998) have proposed smoothing methods based on local least squares and smoothing splines, and recent approaches include a componentwise kernel method (Wu & Chiang, 2000), a componentwise spline method (Chiang *et al.*, 2001) and a method based on local maximum likelihood estimates (Cai *et al.*, 2000). We derive asymptotic distributions for an estimation method that is tailored to this special model.

Since the assumptions on $\tilde{Y}$ and $\tilde{X}$ are symmetric, with a similar argument as above, regressing $\tilde{X}$ on $\tilde{Y}$ and U leads to a second varying coefficient model

$$\tilde{X}_i = \alpha_0(U_i) + \alpha_1(U_i)\tilde{Y}_i + \phi(U_i)e_{XY,i},$$

where

$$\alpha_0(u) = \eta_0\phi(u), \qquad \alpha_1(u) = \eta_1\frac{\phi(u)}{\psi(u)}. \tag{8}$$

3. Estimation of covariate adjusted correlation and asymptotic distribution

The proposed Cadcor estimate for $\rho_{(Y,X)}$ is

$$r = \text{sign}(\hat{\gamma}_1)\sqrt{\hat{\gamma}_1\hat{\eta}_1 1_{\{\text{sign}(\hat{\gamma}_1)=\text{sign}(\hat{\eta}_1)\}}}, \tag{9}$$

where the estimates of $\hat{\gamma}_1$ and $\hat{\eta}_1$ are obtained after an initial binning step and $1_{\{\text{sign}(\hat{\gamma}_1)=\text{sign}(\hat{\eta}_1)\}}$ denotes the indicator for $\text{sign}(\hat{\gamma}_1) = \text{sign}(\hat{\eta}_1)$. We assume that the covariate U is bounded below and above, $-\infty < a \leq U \leq b < \infty$ for real numbers $a < b$, and divide the interval $[a, b]$ into m equidistant intervals denoted by $B_1, \ldots, B_m$, referred to as bins. Given m, the B_j, $j = 1, \ldots, m$ are fixed, but the number of U_i's falling into B_j is random and is denoted by L_j. For every U_i falling in the jth bin, i.e., $U_i \in B_j$, the corresponding observed variables are $\tilde{X}_i$ and $\tilde{Y}_i$.

After binning the data, we fit linear regressions of $\tilde{Y}_i$ on $\tilde{X}_i$ and $\tilde{X}_i$ on $\tilde{Y}_i$, using the data falling within each bin B_j, $j = 1, \ldots, m$. The least squares estimates of the resulting linear regressions for the data in the jth bin are denoted by $\hat{\beta}_j^T = (\hat{\beta}_{0j}, \hat{\beta}_{1j})^T$ and $\hat{\alpha}_j^T = (\hat{\alpha}_{0j}, \hat{\alpha}_{1j})^T$, corresponding to the linear regressions of $\tilde{Y}_i$ on $\tilde{X}_i$ and of $\tilde{X}_i$ on $\tilde{Y}_i$. The estimators of γ_0, γ_1, η_0 and η_1 are then obtained as weighted averages of the $\hat{\beta}_j$'s and $\hat{\alpha}_j$'s, weighted according to the number of data L_j in the jth bin,

$$\hat{\gamma}_0 = \sum_{j=1}^{m} \frac{L_j}{n} \hat{\beta}_{0j} \qquad \hat{\gamma}_1 = \frac{1}{\bar{\tilde{X}}} \sum_{j=1}^{m} \frac{L_j}{n} \hat{\beta}_{1j} \bar{\tilde{X}}'_j, \tag{10}$$

and

$$\hat{\eta}_0 = \sum_{j=1}^{m} \frac{L_j}{n} \hat{\alpha}_{0j} \qquad \hat{\eta}_1 = \frac{1}{\bar{\tilde{Y}}} \sum_{j=1}^{m} \frac{L_j}{n} \hat{\alpha}_{1j} \bar{\tilde{Y}}'_j, \tag{11}$$

where $\bar{\tilde{X}} = n^{-1} \sum_{i=1}^{n} \tilde{X}_i$, $\bar{\tilde{Y}} = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i$ and $\bar{\tilde{X}}'_j$, $\bar{\tilde{Y}}'_j$ are the averages of the $\tilde{X}_i$ and $\tilde{Y}_i$ falling into the bin B_j respectively, i.e. $\bar{\tilde{X}}'_j = L_j^{-1} \sum_{i=1}^{n} \tilde{X}_i \mathbf{1}_{\{U_i \in B_j\}}$ and $\bar{\tilde{Y}}'_j = L_j^{-1} \sum_{i=1}^{n} \tilde{Y}_i \mathbf{1}_{\{U_i \in B_j\}}$. (Şentürk & Müller, 2003a). These estimates are motivated by $E\{\beta_0(U)\} = \gamma_0$, $E\{\beta_1(U)\tilde{X}\} = \gamma_1 E(\tilde{X})$, $E\{\alpha_0(U)\} = \eta_0$ and $E\{\alpha_1(U)\tilde{Y}\} = \eta_1 E(\tilde{Y})$ (see (7), (8) and the identifiability conditions).

Remark 1*:* An alternative to the equidistant binning adopted in this paper would be nearest neighbor binning, where the number of points falling in each bin is fixed, but the bin boundaries are random. We have chosen to adopt the equidistant binning where bin boundaries are fixed but the number of points falling in each bin is random, as we found it to work quite well in applications. One might expect equidistant binning to be superior in terms of controlling the bias, whereas the nearest neighbor approach might yield smaller variance, depending on the underlying assumptions. This trade-off could be

further investigated in future research.

Remark 2*:* Yet another alternative to the proposed estimation procedure targeting the regression coefficients γ_0, γ_1, η_0 and η_1 would be to use one of the above cited smoothing techniques, such as kernel smoothing based on local least squares or smoothing splines, to estimate the varying coefficient functions evaluated at the original observation points U_i, $i = 1, \ldots, n$, and then to apply an averaging method to estimate the targeted regression coefficients. In the special case of using local polynomial fitting based on local least squares, Zhang *et al.* (2002) were able to derive the asymptotic conditional mean squared error of such an estimator. However, asymptotic distributional results as we provide in this paper are yet to be established for this class of varying coefficient estimators.

Remark 3*:* A possible extension of the proposed confounding setting would be to consider a vector valued confounding variable U. The proposed binning and smoothing techniques could be easily adapted to two or three dimensional cases, but one will encounter the problem of curse of dimensionality for higher dimensions, as the bins will get sparse quickly as dimension increases. To overcome this problem, a topic for future research would be to adopt a single index approach, where the confounding effects are functions of a linear combination of the confounding vector U, say $\nu^T U$, where the single index vector ν needs to be estimated.

We derive the asymptotic distribution of r, the Cadcor estimate (9), as the number of subjects n tends to infinity. As in typical smoothing applications, the number of bins $m = m(n)$ is required to satisfy $m \to \infty$ and $n/(m \log n) \to \infty$ and $m/\sqrt{n} \to \infty$ as $n \to \infty$. We denote convergence in distribution by $\xrightarrow{\mathcal{D}}$ and convergence in probability by $\xrightarrow{p}$.

Theorem 1. *Under the technical conditions* $(C1) - (C6)$ *given in Appendix A, on events* E_n *(defined in (16)-(18) below) with* $P(E_n) \to 1$ *as* $n \to \infty$*, it holds that*

$$\sqrt{n}(r - \rho_{(Y,X)}) \xrightarrow{\mathcal{D}} \mathbb{N}(0, \sigma_r^2),$$

for $\rho_{(Y,X)} \neq 0$*, where the explicit form of* σ_r^2 *is given in Appendix B.*

Remark 4*:* It is of interest to test the hypothesis $H_0 : \rho_{(Y,X)} = 0$, since this case is

excluded by the assumptions of Theorem 1. Equation (5) implies that testing the hypothesis $H_0 : \rho_{(Y,X)} = 0$ is equivalent to testing $H_0 : \gamma_1 = 0$ or testing $H_0 : \eta_1 = 0$. Thus, we propose to test the hypothesis $H_0 : \rho_{(Y,X)} = 0$ via testing $H_0 : \gamma_1 = 0$. For this testing problem, the bootstrap test proposed in Şentürk & Müller (2003a) is available. This test has been shown to attain desirable coverage levels in simulation studies.

4. Estimating the asymptotic variance

From this point on, we will use subscripts n to denote variables that form a triangular array scheme. The observable data are of the form $(U_{ni}, \tilde{X}_{ni}, \tilde{Y}_{ni})$, $i = 1, \ldots, n$, for a sample of size n. Correspondingly, the underlying unobservable variables and errors are $(X_{ni}, Y_{ni}, e_{YX,ni}, e_{XY,ni})$, $i = 1, \ldots, n$. Let $\{(U'_{njk}, \tilde{X}'_{njk}, \tilde{Y}'_{njk}, X'_{njk}, Y'_{njk}, e'_{YX,njk}, e'_{XY,njk})$, $k = 1, \ldots, L_{nj}, r = 1, \ldots, p\} = \{(U_{ni}, \tilde{X}_{ni}, \tilde{Y}_{ni}, X_{ni}, Y_{ni}, e_{YX,ni}, e_{XY,ni}), i = 1, \ldots, n,$ $r = 1, \ldots, p : U_{ni} \in B_{nj}\}$ denote the data for which $U_{ni} \in B_{nj}$, where we refer to $(U'_{njk}, \tilde{X}'_{njk}, \tilde{Y}'_{njk}, X'_{njk}, Y'_{njk}, e'_{YX,njk}, e'_{XY,njk})$ as the kth element in bin B_{nj}. Then we can express the least squares estimates of the linear regressions of the observable data falling in the jth bin B_{nj} as

$$\hat{\beta}_{n1j} = \frac{\sum_{k=1}^{L_{nj}} (\tilde{X}'_{njk} - \bar{\tilde{X}}'_{nj}) \tilde{Y}'_{njk}}{\sum_{k=1}^{L_{nj}} (\tilde{X}'_{njk} - \bar{\tilde{X}}'_{nj})^2}, \qquad \hat{\beta}_{n0j} = \bar{\tilde{Y}}'_{nj} - \hat{\beta}_{n1j} \bar{\tilde{X}}'_{nj}, \tag{12}$$

for the regression of $\tilde{Y}$ on $\tilde{X}$ leading to the parameter estimates $\hat{\gamma}_{n0}$ and $\hat{\gamma}_{nr}$ given in (10), and

$$\hat{\alpha}_{n1j} = \frac{\sum_{k=1}^{L_{nj}} (\tilde{Y}'_{njk} - \bar{\tilde{Y}}'_{nj}) \tilde{X}'_{njk}}{\sum_{k=1}^{L_{nj}} (\tilde{Y}'_{njk} - \bar{\tilde{Y}}'_{nj})^2}, \qquad \hat{\alpha}_{n0j} = \bar{\tilde{X}}'_{nj} - \hat{\alpha}_{n1j} \bar{\tilde{Y}}'_{nj}, \tag{13}$$

for the regression of $\tilde{X}$ on $\tilde{Y}$ leading to the parameter estimates $\hat{\eta}_{n0}$ and $\hat{\eta}_{nr}$ given in (11) respectively. In the above expression, $\bar{\tilde{X}}'_{nj} = L_{nj}^{-1} \sum_{k=1}^{L_{nj}} \tilde{X}'_{njk}$ and $\bar{\tilde{Y}}'_{nj} = L_{nj}^{-1} \sum_{k=1}^{L_{nj}} \tilde{Y}'_{njk}$.

Next we introduce least squares estimates of the linear regressions of the unobservable data falling into B_{nj}, i.e.,

$$\zeta_{n1j} = \frac{\sum_{k=1}^{L_{nj}} (X'_{njk} - \bar{X}'_{nj}) Y'_{njk}}{\sum_{k=1}^{L_{nj}} (X'_{njk} - \bar{X}'_{nj})^2}, \qquad \zeta_{n0j} = \bar{Y}'_{nj} - \zeta_{n1j} \bar{X}'_{nj}, \tag{14}$$

for the regression of Y on X, and

$$\omega_{n1j} = \frac{\sum_{k=1}^{L_{nj}}(Y'_{njk} - \bar{Y}'_{nj})X'_{njk}}{\sum_{k=1}^{L_{nj}}(Y'_{njk} - \bar{Y}'_{nj})^2}, \qquad \omega_{n0j} = \bar{X}'_{nj} - \omega_{n1j}\bar{Y}'_{nj}, \tag{15}$$

for the regression of X on Y. These quantities are not estimable, but will be used in the proof of the main results.

For $\hat{\gamma}_{n0}$, $\hat{\gamma}_{n1}$, $\hat{\eta}_{n0}$ and $\hat{\eta}_{n1}$ given in (10) and (11) to be well defined, the least squares estimates $\hat{\beta}_{n0j}$, $\hat{\beta}_{n1j}$, $\hat{\alpha}_{n0j}$ and $\hat{\alpha}_{n1j}$ given in (12), (13) must exist for each bin B_{nj}. This requires that $s^2_{\tilde{X}'_{nj}} = L_{nj}^{-1}\sum_k \tilde{X}'^2_{njk} - (L_{nj}^{-1}\sum_k \tilde{X}'^2_{njk})^2$ and $s^2_{\tilde{Y}'_{nj}} = L_{nj}^{-1}\sum_k \tilde{Y}'^2_{njk} - (L_{nj}^{-1}\sum_k \tilde{Y}'^2_{njk})^2$ are strictly positive for each B_{nj}. Correspondingly, ζ_{n0j}, ζ_{n1j}, ω_{n0j} and ω_{n1j} given in (14), (15) will exist under the condition that $s^2_{X'_{nj}}$, $s^2_{Y'_{nj}} > 0$, where $s^2_{X'_{nj}}$ and $s^2_{Y'_{nj}}$ are defined similar to $s^2_{\tilde{X}'_{nj}}$ and $s^2_{\tilde{Y}'_{nj}}$. Define the events

$$\begin{aligned}
\tilde{A}_n &= \{\omega \in \Omega : \inf_j s^2_{\tilde{X}'_{nj}} > \kappa, \text{ and } \min_j L_{nj} > 1\}, \\
A_n &= \{\omega \in \Omega : \inf_j s^2_{X'_{nj}} > \kappa, \text{ and } \min_j L_{nj} > 1\}, \qquad (16) \\
\tilde{C}_n &= \{\omega \in \Omega : \inf_j s^2_{\tilde{Y}'_{nj}} > \kappa, \text{ and } \min_j L_{nj} > 1\}, \\
C_n &= \{\omega \in \Omega : \inf_j s^2_{Y'_{nj}} > \kappa, \text{ and } \min_j L_{nj} > 1\}, \qquad (17)
\end{aligned}$$

$$E_n = \tilde{A}_n \cap A_n \cap \tilde{C}_n \cap C_n, \tag{18}$$

where $\kappa = \min\{\varrho_x/2, \inf_j(\phi^2(U'^*_{nj}))\varrho_x/2, \varrho_y/2, \inf_j(\psi^2(U'^*_{nj}))\varrho_y/2\}$, ϱ_x and ϱ_y are as defined in (C5), $U'^*_{nj} = L_{nj}^{-1}\sum_{k=1}^{L_{nj}} U'_{njk}$ is the average of the U's in B_{nj}, and $(\Omega, \mathcal{F}, P)$ is the underlying probability space. The estimates $\hat{\gamma}_{n0}$, $\hat{\gamma}_{n1}$, ζ_{n0j}, ζ_{n1j}, $\hat{\eta}_{n0}$, $\hat{\eta}_{n1}$, ω_{n0j} and ω_{n1j} are well defined on events $\tilde{A}_n$, A_n, $\tilde{C}_n$ and C_n. Generalizing a result in Appendix A.3 of Şentürk & Müller (2003b), we have that $P(E_n) \to 1$ as $n \to \infty$, due to the symmetry between $\tilde{Y}$ and $\tilde{X}$.

Theorem 2. *Under the technical conditions* $(C1) - (C6)$ *given in Appendix A, on event* E_n *(defined in (16)-(18)) with* $P(E_n) \to 1$ *as* $n \to \infty$*, it holds that*

$$\hat{\sigma}^2_{nr} \xrightarrow{p} \sigma^2_r,$$

where the explicit form of $\hat{\sigma}^2_{nr}$ *is given in Appendix B.*

5. Applications and Monte Carlo study

Under the technical conditions given in Appendix A, if $\rho_{(Y,X)} \neq 0$, the asymptotic distribution of the Cadcor estimate r according to Theorem 1 is $\sqrt{n}(r - \rho_{(Y,X)})/\sigma_r \xrightarrow{\mathcal{D}} \mathbb{N}(0,1)$, as $n \to \infty$, where σ_r^2 is as in Appendix B. Using the consistent estimate $\hat{\sigma}_{nr}^2$ of σ_r^2 proposed in Theorem 2, it follows from Slutsky's theorem that $\sqrt{n}(r - \rho_{(Y,X)})/\hat{\sigma}_{nr} \xrightarrow{\mathcal{D}} \mathbb{N}(0,1)$, so that an approximate $(1-\varphi)$ asymptotic confidence interval for $\rho_{(Y,X)}$ has the endpoints

$$r \pm z_{\varphi/2}\frac{\hat{\sigma}_{nr}}{\sqrt{n}}. \tag{19}$$

Here $z_{\varphi/2}$ is the $(1-\varphi/2)$th quantile of the standard Gaussian distribution.

5.1. Application to the Boston house price data

We analyze a subset of the Boston house price data (available at http://lib.stat.cmu.edu) of Harrison & Rubinfeld (1978). These data include the following variables: proportion of population of lower educational status (i.e. proportion of adults without high school education and proportion of male workers classified as laborers) (*%LS*), per capita crime rate by town (crime rate, *CR*) and median value of owner-occupied homes in $1000's (house price, *HP*) for 506 towns around Boston. A goal is to identify the factors affecting the house prices in Boston. However, it is hard to separate the effects of different factors on *HP* since there are confounding variables present. Of interest is the correlation between *HP* and *CR*. While we can simply compute the standard correlation between these variables, a more meaningful question is whether these variables are still correlated after adjusting for *%LS*, since we may reasonably expect *%LS* to influence the relationship between *HP* and *CR*. We therefore estimate the Cadcor using *%LS* as the confounding variable U.

The correlation $\rho_{(Y,X)}$ was estimated by the Cadcor method and the results were compared to estimators obtained without adjustment and with the standard correlation adjustment methods that we have discussed above in Section 1, namely using estimates of $\rho_{(\tilde{e}_{\hat{Y}U},\tilde{e}_{\hat{X}U})}$ and $\rho_{(e_{\hat{Y}U},e_{\hat{X}U})}$. The estimators and approximate 95% asymptotic confidence intervals for $\rho_{(Y,X)}$ for these four methods are displayed in Table 1. For the two standard adjustment methods and the case of no adjustment, approximate confidence intervals

were obtained using Fisher's z-transformation. Before forming confidence intervals for the proposed covariate adjusted correlation (5), its significance was tested (see Remark 4 after Theorem 1). It was found to be significant at the 5% level (p-value= 0.048). The approximate confidence interval for the Cadcor (19) was obtained using the variance estimate $\hat{\sigma}^2_{nr}$ (see Theorem 2). The scatter-plots of the raw estimates $(\hat{\beta}_{nr1}, \ldots, \hat{\beta}_{nrm})$ (12) and $(\hat{\alpha}_{nr1}, \ldots, \hat{\alpha}_{nrm})$ (13) vs. the midpoints of the bins $(B_{n1}, \ldots, B_{nm})$ are shown in Figure 2 for intercepts ($r = 0$) and slopes ($r = 1$). The scatter-plots of the raw correlations within each bin $(\hat{r}_{n1}, \ldots, \hat{r}_{nm})$ vs. the midpoints of the bins $(B_{n1}, \ldots, B_{nm})$, where $\hat{r}_{nj}$ is defined as

$$\hat{r}_{nj} = \text{sign}(\hat{\beta}_{n1j})\sqrt{\hat{\beta}_{n1j}\hat{\alpha}_{n1j}1_{\{\text{sign}(\hat{\beta}_{n1j})=\text{sign}(\hat{\alpha}_{n1j})\}}}, \tag{20}$$

are shown in Figure 3, along with scatter-plots of the variables $(\%LS, HP, CR)$.

FIGURE 2 ABOUT HERE

The implementation of the binning includes the merging of sparsely populated bins. Bin widths were chosen aiming at a number of at least three points in each bin, and bins with less than three points were merged with neighboring bins. For this example with $n = 506$, the average number of points per bin was 18, yielding a total of 24 bins after merging.

TABLE 1 ABOUT HERE

The unadjusted correlation $\rho_{(\tilde{Y},\tilde{X})}$ for *HP* and *CR* was found to be -0.3880. When adjusted for *%LS* with Cadcor, the adjusted correlation was found to be -0.2201. This implies that the proportion of population of lower educational status explains a significant amount of the negative correlation between median house price and crime rate. The estimate of nonparametric partial correlation $\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$ was -0.1706, which came closest to the Cadcor estimate, even though the approximate 95% confidence interval based on this estimate did not contain the Cadcor of -0.2201. When adjusting for *%LS* with partial correlation ($\hat{\rho}_{(e_{\tilde{Y}U},e_{\tilde{X}U})} = -0.0868$), *HP* and *CR* were found to be no longer correlated. The reason for the relative poor performance of the partial correlation estimate might be its inability to reflect the nonlinear nature of the relationship between *HP* and *%LS*, as demonstrated in Figure 3.

FIGURE 3 ABOUT HERE

From Figure 3 bottom right panel, the correlation between *HP* and *CR* is seen to be positive when *%LS* is between 0 and 10, and to become negative for larger *%LS* values. The same phenomenon can also be observed from the positive slopes of *HP* and *CR* in Figure 2, top and bottom right panels. This positive correlation between house prices and crime rate for relatively high status neighborhoods seems counter-intuitive at first. However, the reason for this positive correlation is the existence of high educational status neighborhoods close to central Boston where high house prices and crime rate occur simultaneously. The expected effect of increasing crime rate on declining house prices seems to be only observed for lower educational status neighborhoods in Boston.

5.2. Monte Carlo simulation

The confounding covariate U was simulated from Uniform$(1, 7)$. The underlying unobserved variables $(X, Y)^T$ were simulated from a bivariate normal distribution with mean vector $(2, 3)^T$, $\sigma_X^2 = 0.3$, $\sigma_Y^2 = 0.4$ and $\rho_{(X,Y)} = 0.5$. The distorting functions were chosen as $\psi(U) = (U + 1)/5$ and $\phi(U) = (-2U^2 + 4U + 90)/68$, satisfying the identifiability conditions. We conducted 1000 Monte Carlo runs with sample sizes 100, 400 and 1600. For each run, 95% asymptotic confidence intervals were constructed by plugging estimates $\hat{\sigma}_{nr}^2$, $r = 0, \ldots, p$, given in Theorem 2, into (19). The estimated non-coverage rates in percent and mean interval lengths for these confidence intervals were (8.06, 5.50, 5.30) and (0.44, 0.16, 0.08) respectively for sample sizes $n = (100, 400, 1600)$. The estimated non-coverage level is getting very close to the target value 0.05, as the sample size increases, and the estimated interval lengths are decreasing.

For the sample size of 400, biases for non-adjustment ($\rho_{(\tilde{Y},\tilde{X})}$) and the two standard correlation adjustment methods, nonparametric partial correlation ($\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$) and partial correlation ($\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$), were also estimated, and found to be 0.9557, 0.1101 and 0.1106 respectively, for the three methods. The bias of the Cadcor algorithm was 0.0080, and thus negligible in comparison to the other methods.

We have also carried out simulations to study the effects of different choices of m, the total number of bins, on the mean square error of the Cadcor estimates. Under the rate conditions on m given in Section 3, the estimates are found to be quite robust regarding

different choices of m; we advocate a choice such that all or the vast majority of bins include at least three observations. If there are a few bins that have less than this minimum number, they will be merged with neighboring bins in a second bin-merging step. The average number of points per bin were $5, 16$ and 32 for sample sizes $n = 100, 400$ and 1600, respectively.

6. Proofs of the main results

Defining $\delta_{n0jk} = \psi(U'_{njk}) - \psi(U'^{*}_{nj})$ and $\delta_{n1jk} = \phi(U'_{njk}) - \phi(U'^{*}_{nj})$ for $1 \le k \le L_j$, where $U'^{*}_{nj} = L_{nj}^{-1} \sum_{k=1}^{L_{nj}} U'_{njk}$ is the average of the U's in B_{nj}, we obtain the following results, by Taylor expansions and boundedness considerations: (a.) $\sup_{k,j} |U'_{njk} - U'^{*}_{nj}| \le (b-a)/m$; (b.) $\sup_{k,j} |\delta_{nrjk}| = O(m^{-1})$, for $r = 0, 1$.

Proof of Theorem 1. We show

$$\sqrt{n} \begin{pmatrix} \sum_{j=1}^{m} \frac{L_{nj}}{n} \hat{\beta}_{n1j} \bar{\tilde{X}}'_{nj} & - & \gamma_1 E(X) \\ \sum_{j=1}^{m} \frac{L_{nj}}{n} \bar{\tilde{X}}'_{nj} & - & E(X) \\ \sum_{j=1}^{m} \frac{L_{nj}}{n} \hat{\alpha}_{n1j} \bar{\tilde{Y}}'_{nj} & - & \eta_1 E(Y) \\ \sum_{j=1}^{m} \frac{L_{nj}}{n} \bar{\tilde{Y}}'_{nj} & - & E(Y) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathbb{N}_4(\underline{0}, \Sigma), \tag{21}$$

where $\Sigma = \{(\sigma_{ij})\}$ is a 4×4 matrix with elements defined in Theorem 1. The asymptotic normality of $\sqrt{n}(\sqrt{\hat{\gamma}_{n1}\hat{\eta}_{n1}} - |\rho_{(Y,X)}|)$ will follow from this with a simple application of the δ-method when γ_1 and η_1 or equivalently $\rho_{(Y,X)}$ are different from zero, using the transformation $g(x_1, x_2, x_3, x_4)^T = \sqrt{(x_1 x_3)/(x_2 x_4)}$. The asymptotic normality of $\sqrt{n}(r_n - \rho_{(Y,X)})$ and thus Theorem 1 will then follow by Slutsky's theorem, since the consistency of $\text{sign}(\hat{\gamma}_{n1})$ and $1_{\{\text{sign}(\hat{\gamma}_1) = \text{sign}(\hat{\eta}_1)\}}$ for $\text{sign}(\rho_{(Y,X)})$ and 1 respectively follow from the consistency of $\hat{\gamma}_{n1}$ and $\hat{\eta}_{n1}$ for γ_1 and η_1 respectively, when γ_1 and η_1 are different from zero, shown in Şentürk & Müller (2003a).

By the Cramer-Wald device it is enough to show the asymptotic normality of

$$\begin{aligned} &\sqrt{n}\Bigg[a\Bigg\{\sum_{j=1}^{m} L_{nj} n^{-1} \hat{\beta}_{n1j} \bar{\tilde{X}}'_{nj} - \gamma_1 E(X)\Bigg\} + b\Bigg\{\sum_{j=1}^{m} L_{nj} n^{-1} \bar{\tilde{X}}'_{nj} - E(X)\Bigg\} \\ &+c\Bigg\{\sum_{j=1}^{m} L_{nj} n^{-1} \hat{\alpha}_{n1j} \bar{\tilde{Y}}'_{nj} - \eta_1 E(Y)\Bigg\} + d\Bigg\{\sum_{j=1}^{m} L_{nj} n^{-1} \bar{\tilde{Y}}'_{nj} - E(Y)\Bigg\}\Bigg] \end{aligned} \tag{22}$$

for real a, b, c, d, and (21) will follow. Let

$$\begin{aligned} S_{nt} &= \sum_{i=1}^{t} Z_{ni} = \sum_{\substack{i=1 \\ j,k}}^{t} \Bigg[a\frac{\gamma_1}{\sqrt{n}}\{\psi(U_{ni})X_{ni} - E(X)\} + \frac{a}{\sqrt{n}}\bar{X}'_{nj(i)}\psi(U_{ni})\frac{(X_{ni} - \bar{X}'_{nj(i)})}{s^2_{X'_{nj(i)}}}e_{YX,ni} \\ &+ \frac{b}{\sqrt{n}}\{\phi(U_{ni})X_{ni} - E(X)\} + c\frac{\eta_1}{\sqrt{n}}\{\phi(U_{ni})Y_{ni} - E(Y)\} \\ &+ \frac{c}{\sqrt{n}}\bar{Y}'_{nj(i)}\phi(U_{ni})\frac{(Y_{ni} - \bar{Y}'_{nj(i)})}{s^2_{Y'_{nj(i)}}}e_{XY,ni} + \frac{d}{\sqrt{n}}\{\psi(U_{ni})Y_{ni} - E(Y)\}\Bigg], \end{aligned}$$

where $\bar{X}'_{nj(i)} = L_{nj}^{-1}\sum_{k=1}^{L_{nj(i)}} X'_{nj(i)k}$ and $s^2_{X'_{nj(i)}} = L_{nj}^{-1}\sum_{k=1}^{L_{nj(i)}} X'^2_{nj(i)k} - \bar{X}'^2_{nj(i)}$ represents the sample mean and variance of the X's falling in the same bin as X_{ni}; and $\bar{Y}'_{nj(i)}$ and $s^2_{Y'_{nj(i)}}$ are defined similarly.

It is shown in Şentürk & Müller (2003b) (equation (17)) that

$$\sup_j \left| \begin{array}{ccc} \hat{\beta}_{n0j} & - & \psi(U'^*_{nj})\zeta_{n0j} \\ \hat{\beta}_{n1j} & - & \{\psi(U'^*_{nj})/\phi(U'^*_{nj})\}\zeta_{n1j} \end{array} \right| = O_p(m^{-1})\mathbf{1}_{2\times 1}, \tag{23}$$

where ζ_{n0j} and ζ_{n1j} are as defined in (14). This result implies that

$$\sup_j \left| \begin{array}{ccc} \hat{\alpha}_{n0j} & - & \phi(U'^*_{nj})\omega_{n0j} \\ \hat{\alpha}_{n1j} & - & \{\phi(U'^*_{nj})/\psi(U'^*_{nj})\}\omega_{n1j} \end{array} \right| = O_p(m^{-1})\mathbf{1}_{2\times 1}, \tag{24}$$

where ω_{n0j} and ω_{n1j} are as defined in (15) and $\mathbf{1}_{2\times 1}$ is a 2×1 vector of ones, since all assumptions for $\tilde{Y}$ and $\tilde{X}$ are symmetric under the correlation setting. It also follows from Lemma 4 (a,b) of Şentürk & Müller (2003b) that $\sup_j |\bar{X}'_{nj} - E(X)| = o_p(1)$, $\sup_j |s^2_{X'_{nj}} - \mathrm{var}(X)| = o_p(1)$, $\sup_j |\bar{e}'_{YX,nj}| = o_p(1)$, and $\sup_j |L_{nj}^{-1}\sum_k X'_{njk}e'_{YX,njk}| = o_p(1)$. This result can analogously be extended to $\sup_j |\bar{Y}'_{nj} - E(Y)| = o_p(1)$, $\sup_j |s^2_{Y'_{nj}} - \mathrm{var}(Y)| = o_p(1)$, $\sup_j |\bar{e}'_{XY,nj}| = o_p(1)$, and $\sup_j |L_{nj}^{-1}\sum_k Y'_{njk}e'_{XY,njk}| = o_p(1)$. It follows from (23), (24), Lemma 4 (a,b) of Şentürk & Müller (2003b) and property (b) that the expression in (22) equals $S_{nn} + O_p(m^{-1}\sqrt{n})$. Since $O_p(m^{-1}\sqrt{n})$ is asymptotically negligible, the expression in (22) is asymptotically equivalent to S_{nn}.

Let $\mathcal{F}_{nt}$ be the σ-field generated by $\{e_{YX,n1}, \ldots, e_{YX,nt}, e_{XY,n1}, \ldots, e_{XY,nt}, U_{n1}, \ldots, U_{nt}, L_{nj(1)}, \ldots, L_{nj(t)}, X'_{nj(1)}, \ldots, X'_{nj(t)}, Y'_{nj(1)}, \ldots, Y'_{nj(t)}\}$. Then it is easy to check that $\{S_{nt} = \sum_{i=1}^{t} Z_{ni}, \mathcal{F}_{nt}, 1 \le t \le n\}$ is a mean zero martingale for $n \ge 1$. Since the σ-fields

are nested, that is, $\mathcal{F}_{nt} \subseteq \mathcal{F}_{n,t+1}$ for all $t \leq n$, using Lemma 1 given in Appendix C, $S_{nn} \xrightarrow{\mathcal{D}} \mathbb{N}(0, (a,b,c,d)\Sigma(a,b,c,d)^T = \sigma_Z^2)$ in distribution (McLeish, 1974, Theorem 2.3 and subsequent discussion], and Theorem 1 follows.

Proof of Theorem 2. The consistency of $\hat{\sigma}_{n11}$, $\hat{\sigma}_{n12}$ and $\hat{\sigma}_{n22}$ for σ_{11}, σ_{12} and σ_{22} are given in Theorem 2 of Şentürk & Müller (2003b). The consistency of $\hat{\sigma}_{n33}$, $\hat{\sigma}_{n34}$ and $\hat{\sigma}_{n44}$ for σ_{33}, σ_{34} and σ_{44} follows from this result, since the assumptions regarding $\tilde{Y}$ and $\tilde{X}$ are symmetric.

The symmetry of $\tilde{Y}$ and $\tilde{X}$ allows to extend equation (22) of Şentürk & Müller (2003b) to

$$\sup_j \left| \begin{array}{lcl} \hat{\alpha}_{n0j} & - & \phi(U'^*_{nj})\eta_0 \\ \hat{\alpha}_{n1j} & - & \{\phi(U'^*_{nj})/\psi(U'^*_{nj})\}\eta_1 \end{array} \right| = o_p(1)\mathbf{1}_{2\times 1}. \tag{25}$$

By Lemma 4 (a,b) and (22) of Şentürk & Müller (2003b), property (b.), Law of Large Numbers and boundedness considerations,

$$\begin{aligned} \hat{\sigma}_{n14} &= n^{-1}\sum_j \psi^2(U'^*_{nj})\bar{X}'_{nj}s^{-2}_{X'_{nj}}\left(\sum_k X'_{njk}Y'_{njk}e'_{YX,njk} - \bar{X}'_{nj}\sum_k Y'_{njk}e'_{YX,njk}\right) \\ &+ n^{-1}\gamma_1\sum_j \psi^2(U'^*_{nj})\sum_k X'_{njk}Y'_{njk} - \gamma_1 E(X)E(Y) + o_p(1) = \sigma_{14} + o_p(1). \end{aligned}$$

With similar arguments, using Lemma 4 (a,b) of Şentürk & Müller (2003b), (25), property (b.), Law of Large Numbers and boundedness considerations, it can be shown that $\hat{\sigma}_{n23} = \sigma_{23} + o_p(1)$. It also holds by the Law of Large Numbers that $\hat{\sigma}_{n24} = \sigma_{24} + o_p(1)$. Using this result, Lemma 4 (a,b) and (22) of Şentürk & Müller (2003b), (25), property (b.), Law of Large Numbers and boundedness considerations, it holds that

$$\begin{aligned} \hat{\sigma}_{n13} &= \eta_1\frac{1}{n}\sum_j \psi(U'^*_{nj})\phi(U'^*_{nj})\frac{\bar{X}'_{nj}}{s^2_{X'_{nj}}}\left(\sum_k X'_{njk}Y'_{njk}e'_{YX,njk} - \bar{X}'_{nj}\sum_k Y'_{njk}e'_{YX,njk}\right) \\ &+ \gamma_1\frac{1}{n}\sum_j \psi(U'^*_{nj})\phi(U'^*_{nj})\frac{\bar{Y}'_{nj}}{s^2_{Y'_{nj}}}\left(\sum_k X'_{njk}Y'_{njk}e'_{XY,njk} - \bar{Y}'_{nj}\sum_k X'_{njk}e'_{XY,njk}\right) \\ &+ \frac{1}{n}\sum_j \psi(U'^*_{nj})\phi(U'^*_{nj})\frac{\bar{X}'_{nj}\bar{Y}'_{nj}}{s^2_{X'_{nj}}s^2_{Y'_{nj}}}\left(\sum_k X'_{njk}Y'_{njk}e'_{YX,njk}e'_{XY,njk} - \bar{X}'_{nj}\sum_k Y'_{njk}e'_{YX,njk}e'_{XY,njk}\right) \\ &- \frac{1}{n}\sum_j \psi(U'^*_{nj})\phi(U'^*_{nj})\frac{\bar{X}'_{nj}\bar{Y}'^2_{nj}}{s^2_{X'_{nj}}s^2_{Y'_{nj}}}\left(\sum_k X'_{njk}e'_{YX,njk}e'_{XY,njk} - \bar{X}'_{nj}\sum_k e'_{YX,njk}e'_{XY,njk}\right) \end{aligned}$$

$$+\gamma_1\eta_1\mathrm{cov}(\tilde{X},\tilde{Y}) + o_p(1) = \sigma_{13} + o_p(1).$$

Acknowledgements

The authors wish to express their gratitude to the two referees, the associate editor and the editor, whose remarks greatly improved the paper. This research was supported in part by NSF grants.

References

Cai, Z., Fan, J. & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Chiang, C., Rice, J. A. & Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96**, 605-617.

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and demand for clean air. *Journal of Environmental Economics & Management* **5**, 81-102.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B* **55**, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809-822.

Lai, T. L., Robbins, H. & Wei, C. Z. (1979). Strong consistency of least-squares estimates in multiple regression 2. *J. Multivariate Anal.* **9**, 343-361.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* **2**, 620-628.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, heredity and panmixia. *Karl Pearson's early statistical papers.* Cambridge University Press (1956), London. 113-178.

Şentürk, D. & Müller, H. G. (2005a). Covariate adjusted regression. *In press in Biometrika.*

Şentürk, D. & Müller, H. G. (2005b). Inference for covariate adjusted regression via varying coefficient models. *Technical Report.*

Wu, C. O. & Chiang, C. T. (2000). Kernel smoothing in varying coefficient models with

longitudinal dependent variable. *Statistica Sinica* **10**, 433-456.

Zhang, W., Lee, S.-Y. & Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *J. Multivariate Anal.* **82**, 166-188.

Damla Şentürk, Department of Statistics, Pennsylvania State University,
University Park, PA 16802, USA, E-mail:dsenturk@stat.psu.edu

Appendix

Appendix A: Technical conditions
We introduce some technical conditions:

(C1) The covariate U is bounded below and above, $-\infty < a \leq U \leq b < \infty$ for real numbers $a < b$. The density $f(u)$ of U satisfies $\inf_{a\leq u\leq b} f(u) > c_1 > 0$, $\sup_{a\leq u\leq b} f(u) < c_2 < \infty$ for real c_1, c_2, and is uniformly Lipschitz continuous, i.e., there exists a real number M such that $\sup_{a\leq u\leq b} |f(u+c) - f(u)| \leq M|c|$ for any real number c.

(C2) The variables (e_{YX}, U, X) and (e_{XY}, U, Y) are mutually independent, where e_{YX}, e_{XY} are as in (6).

(C3) For both variables X and Y, $\sup_{1\leq i\leq n} |X_{ni}| \leq B_1$, $\sup_{1\leq i\leq n} |Y_{ni}| \leq B_2$ for some bounds $B_1, B_2 \in \mathbb{R}$; and $E(X) \neq 0$, $E(Y) \neq 0$.

(C4) Contamination functions $\psi(\cdot)$ and $\phi(\cdot)$ are twice continuously differentiable, satisfying

$$E\psi(U) = 1, \quad E\phi(U) = 1, \quad \phi(\cdot) > 0, \quad \psi(\cdot) > 0.$$

(C5) Variances of X and Y, σ_X^2 and σ_Y^2 are strictly positive, i.e. $\sigma_X^2 > \varrho_x > 0$, $\sigma_Y^2 > \varrho_y > 0$.

(C6) The functions $h_1(u) = \int x g_1(x,u)dx$, $h_2(u) = \int x g_2(x,u)dx$, $h_3(u) = \int y g_3(y,u)dy$ and $h_4(u) = \int y g_4(y,u)dy$ are uniformly Lipschitz continuous, where $g_1(\cdot,\cdot)$, $g_2(\cdot,\cdot)$, $g_3(\cdot,\cdot)$ and $g_4(\cdot,\cdot)$ are the joint density functions of (X,U), (Xe_{YX},U), (Y,U) and (Ye_{XY},U), respectively.

Under these assumptions, the regressions of $\tilde{Y}$ on $\tilde{X}$ and of $\tilde{X}$ on $\tilde{Y}$ both satisfy the conditions given in Şentürk & Müller (2003b). Bounded covariates are standard in asymptotic theory for least squares regression, as are conditions (C2) and (C5) (see Lai

et al., 1979). The identifiability conditions stated in (C4) are equivalent to

$$E(\tilde{Y}|X) = E(Y|X), \qquad E(\tilde{X}|Y) = E(X|Y).$$

This means that the confounding of Y and X by U does not change the mean regression functions. Conditions (C1) and (C6) are needed for the proof of Lemma 4 of Şentürk & Müller (2003b).

Appendix B: Explicit forms for the asymptotic variance of r and its estimate

The asymptotic variance of r is defined as follows

$$\begin{aligned}
\sigma_r^2 &= \frac{\sigma_{11}\eta_1}{4\{E(X)\}^2\gamma_1} - \frac{\sigma_{12}\eta_1}{2\{E(X)\}^2} + \frac{\sigma_{22}\gamma_1\eta_1}{4\{E(X)\}^2} + \frac{\sigma_{13} - \sigma_{14}\eta_1 - \sigma_{23}\gamma_1 + \sigma_{24}\eta_1\gamma_1}{2E(X)E(Y)} \\
&+ \frac{\sigma_{33}\gamma_1}{4\{E(Y)\}^2\eta_1} - \frac{\sigma_{34}\gamma_1}{2\{E(Y)\}^2} + \frac{\sigma_{44}\eta_1\gamma_1}{4\{E(Y)\}^2},
\end{aligned}$$

where

$$\begin{aligned}
\sigma_{11} &= \gamma_1^2[E\{\psi^2(U)\}E(X^2) - \{E(X)\}^2] + \sigma_{YX}^2\frac{\{E(X)\}^2E\{\psi^2(U)\}}{\mathrm{var}(X)}, \\
\sigma_{12} &= \gamma_1[E\{\psi(U)\psi(U)\}E(X^2) - \{E(X)\}^2], \\
\sigma_{22} &= \mathrm{var}(\tilde{X}), \\
\sigma_{13} &= \gamma_1\eta_1\mathrm{cov}(\tilde{X},\tilde{Y}) + \frac{E(X)E(Y)}{\mathrm{var}(X)\mathrm{var}(Y)}E\{\psi(U)\phi(U)\}\mathrm{cov}(X, Ye_{YX}e_{XY}) \\
&+ \gamma_1\frac{E(Y)}{\mathrm{var}(Y)}E\{\psi(U)\phi(U)\}\mathrm{cov}(Y, Xe_{XY}) + \eta_1\frac{E(X)}{\mathrm{var}(X)}E\{\psi(U)\phi(U)\}\mathrm{cov}(X, Ye_{YX}) \\
&- \frac{E(X)\{E(Y)\}^2}{\mathrm{var}(X)\mathrm{var}(Y)}E\{\psi(U)\phi(U)\}\mathrm{cov}(X, e_{YX}e_{XY}), \\
\sigma_{14} &= \gamma_1[E\{\psi^2(U)\}E(XY) - E(X)E(Y)] + \frac{E(X)}{\mathrm{var}(X)}E\{\psi^2(U)\}\mathrm{cov}(X, Ye_{YX}), \\
\sigma_{23} &= \eta_1[E\{\phi^2(U)\}E(XY) - E(X)E(Y)] + \frac{E(Y)}{\mathrm{var}(Y)}E\{\phi^2(U)\}\mathrm{cov}(Y, Xe_{XY}), \\
\sigma_{24} &= \mathrm{cov}(\tilde{X},\tilde{Y}), \\
\sigma_{33} &= \eta_1^2[E\{\phi^2(U)\}E(Y^2) - \{E(Y)\}^2] + \sigma_{XY}^2\frac{\{E(Y)\}^2E\{\phi^2(U)\}}{\mathrm{var}(Y)}, \\
\sigma_{34} &= \eta_1[E\{\phi(U)\psi(U)\}E(Y^2) - \{E(Y)\}^2], \\
\sigma_{44} &= \mathrm{var}(\tilde{Y}).
\end{aligned}$$

The consistent estimate of σ_r^2 is

$$\begin{aligned}
\hat{\sigma}_{nr}^2 &= \frac{\hat{\sigma}_{n11}\hat{\eta}_{n1}}{4\bar{\tilde{X}}_n^2\hat{\gamma}_{n1}} - \frac{\hat{\sigma}_{n12}\hat{\eta}_{n1}}{2\bar{\tilde{X}}_n^2} + \frac{\hat{\sigma}_{n22}\hat{\gamma}_{n1}\hat{\eta}_{n1}}{4\bar{\tilde{X}}_n^2} + \frac{\hat{\sigma}_{n13} - \hat{\sigma}_{n14}\hat{\eta}_{n1} - \hat{\sigma}_{n23}\hat{\gamma}_{n1} + \hat{\sigma}_{n24}\hat{\eta}_{n1}\hat{\gamma}_{n1}}{2\bar{\tilde{X}}_n\bar{\tilde{Y}}_n} \\
&+ \frac{\hat{\sigma}_{n33}\hat{\gamma}_{n1}}{4\bar{\tilde{Y}}_n^2\hat{\eta}_{n1}} - \frac{\hat{\sigma}_{n34}\hat{\gamma}_{n1}}{2\bar{\tilde{Y}}_n^2} + \frac{\hat{\sigma}_{n44}\hat{\eta}_{n1}\hat{\gamma}_{n1}}{4\bar{\tilde{Y}}_n^2},
\end{aligned}$$

where

$$
\begin{aligned}
\hat{\sigma}_{n11} &= \frac{1}{n}\sum_j \hat{\beta}_{n1j}^2 \sum_k \tilde{X}_{njk}'^2 - \hat{\gamma}_{n1}^2 \bar{\tilde{X}}_n + \left(n^{-1}\sum_j\sum_k \hat{\tilde{e}}_{YX,njk}'^2\right)\left(n^{-1}\sum_j L_{nj}\bar{\tilde{X}}_{nj}'^2 s_{\tilde{X}_{nj}'}^{-2}\right),\\
\hat{\sigma}_{n12} &= n^{-1}\sum_j \hat{\beta}_{n1j}\sum_k \tilde{X}_{njk}'^2 - \hat{\gamma}_{n1}\bar{\tilde{X}}_n^2,\\
\hat{\sigma}_{n22} &= s_{\tilde{X}}^2,\\
\hat{\sigma}_{n13} &= \hat{\gamma}_{n1}\hat{\eta}_{n1}\hat{\mathrm{cov}}(\tilde{X},\tilde{Y}) + n^{-1}\sum_j L_{nj}\bar{\tilde{X}}_{nj}'\bar{\tilde{Y}}_{nj}' s_{\tilde{X}_{nj}'}^{-2} s_{\tilde{Y}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{Y}_{nj}'\tilde{e}_{YX,nj}'\tilde{e}_{XY,nj}')\\
&+ n^{-1}\sum_j L_{nj}\hat{\beta}_{n1j}\bar{\tilde{Y}}_{nj}' s_{\tilde{Y}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{Y}_{nj}',\tilde{X}_{nj}'\tilde{e}_{XY,nj}')\\
&+ n^{-1}\sum_j L_{nj}\hat{\alpha}_{n1j}\bar{\tilde{X}}_{nj}' s_{\tilde{X}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{Y}_{nj}'\tilde{e}_{YX,nj}')\\
&- n^{-1}\sum_j L_{nj}\bar{\tilde{X}}_{nj}'\bar{\tilde{Y}}_{nj}'^2 s_{\tilde{X}_{nj}'}^{-2} s_{\tilde{Y}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{e}_{YX,nj}'\tilde{e}_{XY,nj}'),\\
\hat{\sigma}_{n14} &= n^{-1}\sum_j \hat{\beta}_{n1j}\sum_k \tilde{X}_{njk}'\tilde{Y}_{njk}' - \hat{\gamma}_{n1}\bar{\tilde{X}}_n\bar{\tilde{Y}}_n + n^{-1}\sum_j L_{nj}\bar{\tilde{X}}_{nj}' s_{\tilde{X}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{Y}_{nj}'\tilde{e}_{YX,nj}'),\\
\hat{\sigma}_{n23} &= n^{-1}\sum_j \hat{\alpha}_{n1j}\sum_k \tilde{X}_{njk}'\tilde{Y}_{njk}' - \hat{\eta}_{n1}\bar{\tilde{X}}_n\bar{\tilde{Y}}_n + n^{-1}\sum_j L_{nj}\bar{\tilde{Y}}_{nj}' s_{\tilde{Y}_{nj}'}^{-2}\hat{\mathrm{cov}}(\tilde{Y}_{nj}',\tilde{X}_{nj}'\tilde{e}_{XY,nj}'),\\
\hat{\sigma}_{n24} &= \hat{\mathrm{cov}}(\tilde{X},\tilde{Y}),\\
\hat{\sigma}_{n33} &= n^{-1}\sum_j \hat{\alpha}_{n1j}^2\sum_k \tilde{Y}_{njk}'^2 - \hat{\eta}_{n1}^2\bar{\tilde{Y}}_n + \left(n^{-1}\sum_j\sum_k \hat{\tilde{e}}_{XY,njk}'^2\right)\left(n^{-1}\sum_j L_{nj}\bar{\tilde{Y}}_{nj}'^2 s_{\tilde{Y}_{nj}'}^{-2}\right),\\
\hat{\sigma}_{n34} &= n^{-1}\sum_j \hat{\alpha}_{n1j}\sum_k \tilde{Y}_{njk}'^2 - \hat{\eta}_{n1}\bar{\tilde{Y}}_n^2,\\
\hat{\sigma}_{n44} &= s_{\tilde{Y}}^2,
\end{aligned}
$$

$\hat{\tilde{e}}_{YX,njk}' = \tilde{Y}_{njk}' - \hat{\beta}_{n0j} - \hat{\beta}_{n1j}\tilde{X}_{njk}'$, $\hat{\tilde{e}}_{XY,njk}' = \tilde{X}_{njk}' - \hat{\alpha}_{n0j} - \hat{\alpha}_{n1j}\tilde{Y}_{njk}'$, $\hat{\mathrm{cov}}(\tilde{X},\tilde{Y}) = n^{-1}\sum_j\sum_k \tilde{X}_{njk}'\tilde{Y}_{njk}' - \bar{\tilde{X}}_n\bar{\tilde{Y}}_n$, $\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{Y}_{nj}'\tilde{e}_{YX,nj}') = L_{nj}^{-1}\sum_k \tilde{X}_{njk}'\tilde{Y}_{njk}'\hat{\tilde{e}}_{YX,njk}'$ $-\bar{\tilde{X}}_{nj}' L_{nj}^{-1}\sum_k \tilde{Y}_{njk}'\hat{\tilde{e}}_{YX,njk}'$ and $\hat{\mathrm{cov}}(\tilde{Y}_{nj}',\tilde{X}_{nj}'\tilde{e}_{XY,nj}')$, $\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{Y}_{nj}'\tilde{e}_{YX,nj}'\tilde{e}_{XY,nj}')$, $\hat{\mathrm{cov}}(\tilde{X}_{nj}',\tilde{e}_{YX,nj}'\tilde{e}_{XY,nj}')$ are defined similarly.

Appendix C: Auxiliary results on martingale differences

Lemma 1. *Under the technical conditions (C1)-(C6), on events A_n (16) and C_n (17), the martingale differences Z_{nt} satisfy the conditions*

$$(a.) \qquad \sum_{t=1}^n E\{Z_{nt}^2 I(|Z_{nt}| > \epsilon)\} \to 0 \qquad \text{for all} \quad \epsilon > 0,$$

$$(b.) \qquad \Delta_n^2 = \sum_{t=1}^n Z_{nt}^2 \xrightarrow{p} \sigma_Z^2 \qquad \text{for} \;\; \sigma_Z^2 > 0.$$

Proof. Let $Z_{nt} = w_{nt}v_{nt}$, where $w_{nt} = 1/\sqrt{n}$, and

$$\begin{aligned} v_{nt} &= a\gamma_1\{\psi(U_{nt})X_{nt} - E(X)\} + a\bar{X}'_{nj(t)}\psi(U_{nt})e_{YX,nt}\frac{(X_{nt} - \bar{X}'_{nj(t)})}{s^2_{X'_{nj(t)}}} \\ &+ b\{\phi(U_{nt})X_{nt} - E(X)\} + c\eta_1\{\phi(U_{nt})Y_{nt} - E(Y)\} \\ &+ c\bar{Y}'_{nj(t)}\phi(U_{nt})e_{XY,nt}\frac{(Y_{nt} - \bar{Y}'_{nj(t)})}{s^2_{Y'_{nj(t)}}} + d\{\psi(U_{nt})Y_{nt} - E(Y)\} \end{aligned}$$

with $E(v_{nt}) = 0$. Since $|v_{nt}|$ is bounded uniformly in t on event E_n, it holds for $\epsilon > 0$ that $\sum_{t=1}^n E\{Z_{nt}^2 I(|Z_{nt}| > \epsilon)\} = \sum_{t=1}^n \int x^2 I(|x| > \epsilon) dF_{w_{nt}v_{nt}}(x) \leq \max_t \int x^2 I(|x| > \epsilon/|w_{nt}|) dF_{v_{nt}}(x) \sum_t w_{nt}^2 = \max_t \int x^2 I(|x| > \sqrt{n}\epsilon) dF_{v_{nt}}(x) \to 0$, and Lemma 1 (a.) follows.

The term Δ_n^2 in Lemma 1 (b.) is equal to

$$\begin{aligned} \Delta_n^2 &= a^2\gamma_1^2\left[n^{-1}\sum_t \psi^2(U_{nt})X_{nt}^2 + \{E(X)\}^2 - 2E(X)\left\{n^{-1}\sum_t \psi(U_{nt})X_{nt}\right\}\right] \\ &+ b^2\left[n^{-1}\sum_t \tilde{X}_{nt}^2 + \{E(X)\}^2 - 2E(X)\bar{\tilde{X}}_n\right] \\ &+ c^2\eta_1^2\left[n^{-1}\sum_t \phi^2(U_{nt})Y_{nt}^2 + \{E(Y)\}^2 - 2E(Y)\left\{n^{-1}\sum_t \phi(U_{nt})Y_{nt}\right\}\right] \\ &+ d^2\left[n^{-1}\sum_t \tilde{Y}_{nt}^2 + \{E(Y)\}^2 - 2E(Y)\bar{\tilde{Y}}_n\right] \\ &+ 2ab\gamma_1\left[n^{-1}\sum_t \psi(U_{nt})\phi(U_{nt})X_{nt}^2 - E(X)\left\{n^{-1}\sum_t \psi(U_{nt})X_{nt}\right\} - E(X)\bar{\tilde{X}}_n + \{E(X)\}^2\right] \\ &+ 2ac\gamma_1\eta_1\left[n^{-1}\sum_t \tilde{X}_{nt}\tilde{Y}_{nt} - E(Y)\left\{n^{-1}\sum_t \psi(U_{nt})X_{nt}\right\} - E(X)\left\{n^{-1}\sum_t \phi(U_{nt})Y_{nt}\right\}\right. \\ &+ \left. E(X)E(Y)\right] \\ &+ 2ad\gamma_1\left[n^{-1}\sum_t \psi^2(U_{nt})X_{nt}Y_{nt} - E(Y)\left\{n^{-1}\sum_t \psi(U_{nt})X_{nt}\right\} - E(X)\bar{\tilde{Y}}_n + E(X)E(Y)\right] \\ &+ 2bc\eta_1\left[n^{-1}\sum_t \phi^2(U_{nt})X_{nt}Y_{nt} - E(X)\left\{n^{-1}\sum_t \phi(U_{nt})Y_{nt}\right\} - E(Y)\bar{\tilde{X}}_n + E(X)E(Y)\right] \\ &+ 2bd\left[n^{-1}\sum_t \tilde{X}_{nt}\tilde{Y}_{nt} - E(Y)\bar{\tilde{X}}_n - E(X)\bar{\tilde{Y}}_n + E(X)E(Y)\right] \\ &+ 2cd\eta_1\left[n^{-1}\sum_t \phi(U_{nt})\psi(U_{nt})Y_{nt}^2 - E(Y)\left\{n^{-1}\sum_t \phi(U_{nt})Y_{nt}\right\} - E(Y)\bar{\tilde{Y}}_n + \{E(Y)\}^2\right] \end{aligned}$$

$$
\begin{aligned}
&+ \quad 2\Bigg[n^{-1}\sum_t \psi(U_{nt})\bar{X}'_{nj(t)}\frac{(X_{nt}-\bar{X}'_{nj(t)})}{s^2_{X'_{nj(t)}}}e_{YX,nt}[a^2\gamma_1\{\psi(U_{nt})X_{nt}-E(X)\} \\
&+ \quad ab\{\phi(U_{nt})X_{nt}-E(X)\}-E(Y)(ac\eta_1+ad)]\Bigg] \\
&+ \quad 2\Bigg[n^{-1}\sum_t \phi(U_{nt})\bar{Y}'_{nj(t)}\frac{(Y_{nt}-\bar{Y}'_{nj(t)})}{s^2_{Y'_{nj(t)}}}e_{XY,nt}[c^2\eta_1\{\phi(U_{nt})Y_{nt}-E(Y)\} \\
&+ \quad cd\{\psi(U_{nt})Y_{nt}-E(Y)\}-E(X)(ac\gamma_1+bc)]\Bigg] \\
&+ \quad 2n^{-1}\sum_t \psi(U_{nt})Y_{nt}\bar{X}'_{nj(t)}\frac{(X_{nt}-\bar{X}'_{nj(t)})}{s^2_{X'_{nj(t)}}}e_{YX,nt}\{ac\eta_1\phi(U_{nt})+ad\psi(U_{nt})\} \\
&+ \quad 2n^{-1}\sum_t \phi(U_{nt})X_{nt}\bar{Y}'_{nj(t)}\frac{(Y_{nt}-\bar{Y}'_{nj(t)})}{s^2_{Y'_{nj(t)}}}e_{XY,nt}\{ac\gamma_1\psi(U_{nt})+bc\phi(U_{nt})\} \\
&+ \quad a^2n^{-1}\sum_t \psi^2(U_{nt})\bar{X}'^2_{nj(t)}\frac{(X_{nt}-\bar{X}'_{nj(t)})^2}{s^4_{X'_{nj(t)}}}e^2_{YX,nt} \\
&+ \quad 2acn^{-1}\sum_t \psi(U_{nt})\phi(U_{nt})\bar{Y}'_{nj(t)}\bar{X}'_{nj(t)}\frac{(X_{nt}-\bar{X}'_{nj(t)})(Y_{nt}-\bar{Y}'_{nj(t)})}{s^2_{X'_{nj(t)}}s^2_{Y'_{nj(t)}}}e_{YX,nt}e_{XY,nt} \\
&+ \quad c^2n^{-1}\sum_t \phi^2(U_{nt})\bar{Y}'^2_{nj(t)}\frac{(Y_{nt}-\bar{Y}'_{nj(t)})^2}{s^4_{Y'_{nj(t)}}}e^2_{XY,nt} = T_1+\ldots+T_{17}.
\end{aligned}
$$

It follows from the Law of Large Numbers that

$$
\begin{aligned}
&T_1+\ldots+T_{10} \xrightarrow{p} a^2\gamma_1^2[E\{\psi^2(U)\}E(X^2)-\{E(X)\}^2]+b^2\mathrm{var}(\tilde{X})+d^2\mathrm{var}(\tilde{Y}) \\
&+ \quad c^2\eta_1^2[E\{\phi^2(U)\}E(Y^2)-\{E(Y)\}^2]+2ab\gamma_1[E\{\phi(U)\psi(U)\}E(X^2)-\{E(X)\}^2] \\
&+ \quad (2ac\gamma_1\eta_1+2bd)[E\{\phi(U)\psi(U)\}E(XY)-E(X)E(Y)] \\
&+ \quad 2bc\eta_1[E\{\phi^2(U)\}E(XY)-E(X)E(Y)]+2ad\gamma_1[E\{\psi^2(U)\}E(XY)-E(X)E(Y)] \\
&+ \quad 2cd\eta_1[E\{\phi(U)\psi(U)\}E(Y^2)-\{E(Y)\}^2].
\end{aligned}
$$

On event A, $E(T_{11}|U,X,L_{nj})=0$ and $\mathrm{var}(T_{11}|U,X,L_{nj})=4n^{-2}\sigma^2_{YX}\sum_t\psi^2(U_{nt})\bar{X}'^2_{nj(t)}(X'_{nt}-\bar{X}'_{nj(t)})^2s^{-4}_{X'_{nj(t)}}[a^2\gamma_1\{\psi(U_{nt})X_{nt}-E(X)\}+ab\{\phi(U_{nt})X_{nt}-E(X)\}-E(Y)(ac\eta_1+ad)]^2$ which is $O(n^{-1})$. Thus, $E(T_{11})=0$ and $\mathrm{var}(T_{11})=O(n^{-1})$, implying that $T_{11}=O_p(n^{-1/2})$ on A. Similarly, it can be shown that $T_{12}=O_p(n^{-1/2})$ on C. Using Lemma 4 (a,b) of Şentürk & Müller (2003b) and the Law of Large Numbers, it follows that

$$
\begin{aligned}
T_{13} &\xrightarrow{p} \frac{E(X)}{\mathrm{var}(X)}\mathrm{cov}(X,Ye_{YX})[2ac\eta_1E\{\psi(U)\phi(U)\}+2adE\{\psi^2(U)\}], \\
T_{14} &\xrightarrow{p} \frac{E(Y)}{\mathrm{var}(Y)}\mathrm{cov}(Y,Xe_{XY})[2ac\gamma_1E\{\psi(U)\phi(U)\}+2bcE\{\phi^2(U)\}],
\end{aligned}
$$

$$
\begin{aligned}
T_{15} &\xrightarrow{p} a^2\sigma_{YX}^2\frac{\{E(X)\}^2E\{\psi^2(U)\}}{\mathrm{var}(X)},\\
T_{16} &\xrightarrow{p} 2ac\frac{E(X)E(Y)}{\mathrm{var}(X)\mathrm{var}(Y)}E\{\psi(U)\phi(U)\}\mathrm{cov}(X, Ye_{YX}e_{XY})\\
&- 2ac\frac{E(X)\{E(Y)\}^2}{\mathrm{var}(X)\mathrm{var}(Y)}E\{\psi(U)\phi(U)\}\mathrm{cov}(X, e_{YX}e_{XY}),\\
T_{17} &\xrightarrow{p} c^2\sigma_{XY}^2\frac{\{E(Y)\}^2E\{\phi^2(U)\}}{\mathrm{var}(Y)}.
\end{aligned}
$$

Thus $\Delta_n^2 \xrightarrow{p} \sigma_Z^2 = (a, b, c, d)\Sigma(a, b, c, d)^T$, where $\Sigma_{4\times4}$ is as defined in Theorem 1, and Lemma 1 (b.) follows.

Appendix D: Analysis of ξ_1 defined in (3)

Assuming conditions (C1)-(C6) (see Appendix A), we estimate $\rho_{(Y,X)}$ by a consistent estimate of $\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$, where $\tilde{e}_{\tilde{Y}U}$ and $\tilde{e}_{\tilde{X}U}$ are the errors from the nonparametric regression models $\tilde{Y} = E(\tilde{Y}|U) + \tilde{e}_{\tilde{Y}U}$ and $\tilde{X} = E(\tilde{X}|U) + \tilde{e}_{\tilde{X}U}$, respectively. Thus, $\tilde{e}_{\tilde{Y}U} = \tilde{Y} - E(\tilde{Y}|U) = \psi(U)\{Y - E(Y)\}$ and $\tilde{e}_{\tilde{X}U} = \tilde{X} - E(\tilde{X}|U) = \phi(U)\{X - E(X)\}$. Therefore, using the population equation for correlation,

$$
\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})} = \frac{E(\tilde{e}_{\tilde{Y}U}\tilde{e}_{\tilde{X}U}) - E(\tilde{e}_{\tilde{Y}U})E(\tilde{e}_{\tilde{X}U})}{\sqrt{\mathrm{var}(\tilde{e}_{\tilde{X}U})}\sqrt{\mathrm{var}(\tilde{e}_{\tilde{Y}U})}} = \rho_{(Y,X)}\Delta = \xi_1,
$$

where Δ as defined in Section 1 is equal to $E\{\psi(U)\phi(U)\}/[\sqrt{E\{\psi^2(U)\}}\sqrt{E\{\phi^2(U)\}}]$.

Next, we show that Δ can assume any real value in $(0, 1]$ under suitable conditions. Let $\{\theta_1, \theta_2, \theta_3, \ldots\}$ be an orthogonal basis of the inner-product space $\mathcal{C}[a, b]$, the space of continuous functions on $[a, b]$, using the inner product $< g_1, g_2 >= \int_a^b g_1(u)g_2(u)f(u)du$, where $f(\cdot)$ represents the density function of U and we choose $\theta_1 \equiv 1$. Then ψ and ϕ can be expanded as $\psi = \sum_i \mu_i\theta_i$, and $\phi = \sum_i \vartheta_i\theta_i$, for sets of real numbers μ_i, ϑ_i. The identifiability conditions imply that $\mu_1 = \vartheta_1 = 1$. Assume without loss of generality that for a given set of ϑ_i, $i \geq 2$, $\mu_i = \lambda\vartheta_i$ for an arbitrary $\lambda \geq 0$, and that $\sum_{i\geq2} \vartheta_i^2 = \tau$ for $\tau \geq 0$, i.e. $< \phi, \phi >= \tau + 1$. Hence, $\Delta = (1 + \lambda\tau)/(\sqrt{\tau + 1}\sqrt{\lambda^2\tau + 1})$. For the case of $\lambda = 0$, the value of Δ gets arbitrarily close to 0 as τ increases, and $\Delta = 1$ for $\tau = \lambda = 1$. Thus Δ may assume any real value in the interval $(0, 1]$.

Appendix E: Analysis of ξ_2

Partial correlation of $\tilde{Y}$ and $\tilde{X}$ adjusted for U is equivalent to $\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$, where $e_{\tilde{Y}U}$ and $e_{\tilde{X}U}$ are the errors from the regression models $\tilde{Y} = a_0 + a_1U + e_{\tilde{Y}U}$ and $\tilde{X} = b_0 + b_1U + e_{\tilde{X}U}$, respectively. Assuming $\psi(U) = c_0 + c_1U$ and $\phi(U) = d_0 + d_1U$, for some real numbers

c_0, c_1, d_0, d_1, we can evaluate $e_{\tilde{Y}U}$ and $e_{\tilde{X}U}$, and thus $\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$. Using the population normal equations for regression, we find $a_1 = \{E(\tilde{Y}U) - E(\tilde{Y})E(U)\}/\text{var}(U) = c_1E(Y)$, $a_0 = E(\tilde{Y}) - a_1E(U) = c_0E(Y)$, $b_1 = \{E(\tilde{X}U) - E(\tilde{X})E(U)\}/\text{var}(U) = d_1E(X)$, and $b_0 = E(\tilde{X}) - b_1E(U) = d_0E(X)$. Therefore, $e_{\tilde{Y}U} = \psi(U)\{Y - E(Y)\}$, $e_{\tilde{X}U} = \phi(U)\{X - E(X)\}$, and

$$\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})} = \frac{E(e_{\tilde{Y}U}e_{\tilde{X}U}) - E(e_{\tilde{Y}U})E(e_{\tilde{X}U})}{\sqrt{\text{var}(e_{\tilde{X}U})}\sqrt{\text{var}(e_{\tilde{Y}U})}} = \rho_{(Y,X)}\Delta = \xi_2.$$

Appendix F: Analysis of ξ_3 in (4)

Applying the population equation for correlation and simplifying terms, we find $\rho_{(\tilde{Y},\tilde{X})} = \{E(\tilde{Y}\tilde{X}) - E(\tilde{Y})E(\tilde{X})\}/\{\sqrt{\text{var}(\tilde{X})}\sqrt{\text{var}(\tilde{Y})}\} = \xi_3$. Expanding ψ and ϕ in the same way as in Appendix D, $\xi_3 = [(1+\lambda\tau)E(XY) - E(X)E(Y)]/[\sqrt{E(X^2)(\tau+1) - \{E(X)\}^2}$ $\sqrt{E(Y^2)(\lambda^2\tau+1) - \{E(Y)\}^2}]$. This quantity can assume any real value in $[-1, 1]$, since in the special case of $\tau = 0$, $\xi_3 = \rho_{(Y,X)}$.

Table 1: Estimates and approximate 95% confidence intervals for $\rho_{(HP,CR)}$ adjusted for $\%LS$ in the Boston house price data set with $n = 506$. The first three estimates correspond to non-adjustment ($\rho_{(\tilde{Y},\tilde{X})}$), nonparametric partial correlation ($\rho_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$) and partial correlation ($\rho_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$). The approximate confidence intervals for these three methods were obtained using Fisher's z-transformation. The fourth estimate was obtained from the Cadcor method adjusting for $\%LS$, with the asymptotic intervals (19).

Methods	Lower B.	Estimate	Upper B.
$\hat{\rho}_{(\tilde{Y},\tilde{X})}$	-0.4600	-0.3880	-0.3118
$\hat{\rho}_{(e_{\tilde{Y}U},e_{\tilde{X}U})}$	-0.1720	-0.0868	0.0003
$\hat{\rho}_{(\tilde{e}_{\tilde{Y}U},\tilde{e}_{\tilde{X}U})}$	-0.2150	-0.1706	-0.1260
Cadcor	-0.3113	-0.2201	-0.1289

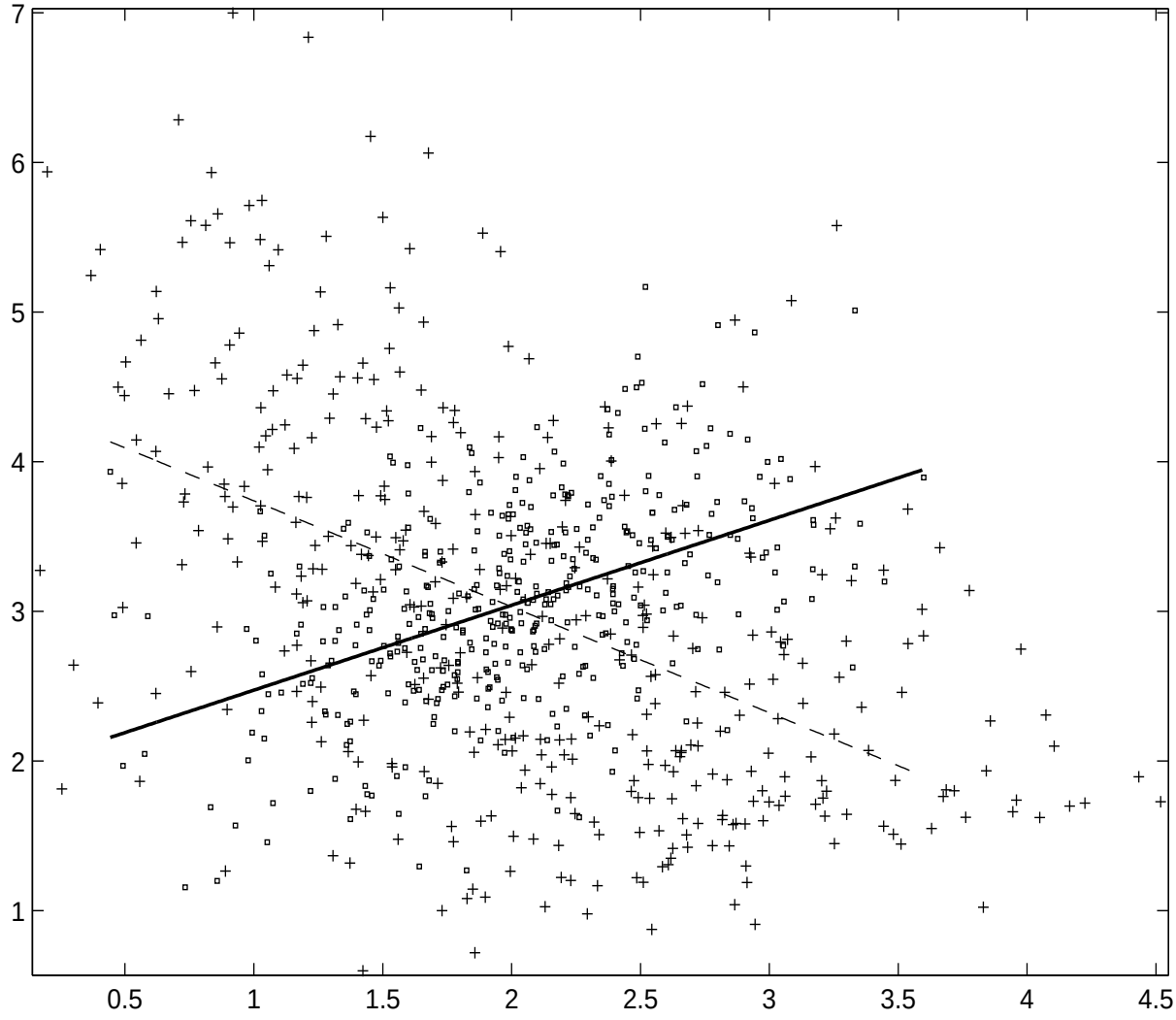

Figure 1: Data (X_i, Y_i) (squares), $i = 1, \ldots, 400$, generated from the underlying bivariate distribution specified in Section 5.2, along with the distorted data $(\tilde{X}_i, \tilde{Y}_i)$ (crosses). Least squares linear lines fitted to distorted data (dashed) (with estimated correlation $\hat{\rho}_{(\tilde{Y},\tilde{X})} = -0.4552$), and to original data (solid) (with estimated correlation $\hat{\rho}_{(Y,X)} = 0.4924$) are also shown.

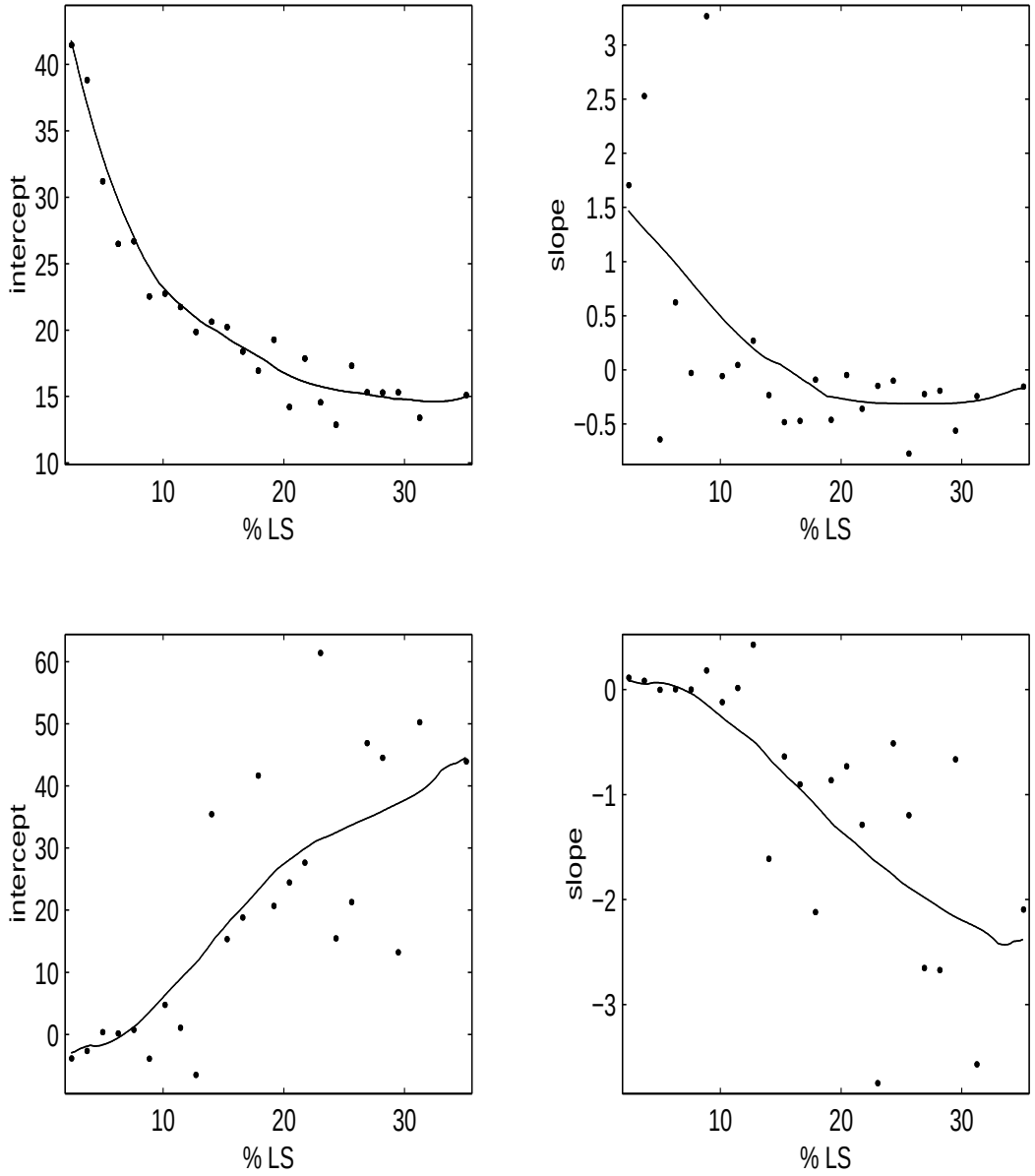

Figure 2: Scatter-plots of the estimated regression coefficients $(\hat{\beta}_{nr1}, \ldots, \hat{\beta}_{nrm})$ (12) of linear regressions house price (HP) versus crime rate (CR) for each bin $(B_{n1}, \ldots, B_{nm})$ for intercepts ($r = 0$, top left panel) and slopes ($r = 1$, top right panel). Similarly, estimated linear regression coefficients $(\hat{\alpha}_{nr1}, \ldots, \hat{\alpha}_{nrm})$ (13) of linear regressions CR vs. HP for each bin $(B_{n1}, \ldots, B_{nm})$ for intercepts ($r = 0$, bottom left panel) and slopes ($r = 1$, bottom right panel). Here $\tilde{Y} = HP$, $\tilde{X} = CR$ and confounding variable $U = \%LS$ (lower educational status). Local linear kernel smooths have been fitted through the scatter-plots using cross validation bandwidth choices of $h = 6, 10, 10, 10$, respectively. Sample size is 506, and the number of bins formed is 24.

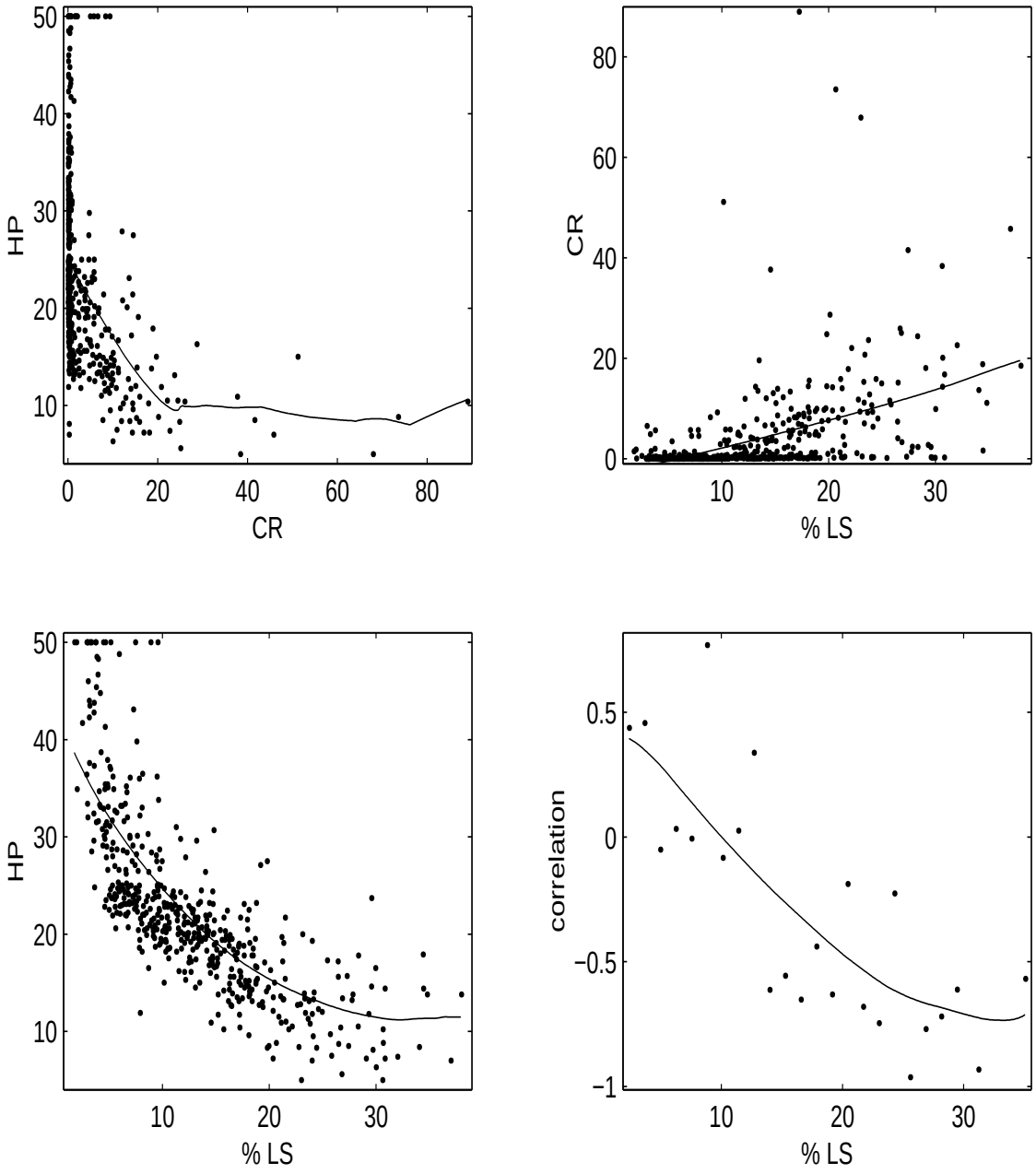

Figure 3: Scatter-plots of the variables HP (house price) vs CR (crime rate) (top left panel), CR vs $\%LS$ (lower educational status)(top right panel) and HP vs $\%LS$ (bottom left panel) for the Boston house price data, along with the scatter-plot of raw correlation estimates $(\hat{r}_{n1}, \ldots, \hat{r}_{nm})$ (20) per each bin $(B_{n1}, \ldots, B_{nm})$ (bottom right panel). Local linear kernel smooths have been fitted through the scatter-plots using cross validation bandwidth choices of $h = 25, 15, 15, 10$, respectively.