I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N
P A P E R

2013/25

To Smooth or Not to Smooth? The Case of Discrete Variables in Nonparametric Regression

LI, D., SIMAR, L. and V. ZELENYUK

To Smooth or Not to Smooth? The Case of Discrete Variables in Nonparametric Regression

Degui Li* Léopold Simar† Valentin Zelenyuk‡

Abstract

In this paper, we consider the nonparametric smoothing technique with both discrete and categorical regressors. In the existing literature, it is generally admitted that it is better to smooth the discrete variables, which is similar to the smoothing technique for continuous regressors but using discrete kernels. However, as we explain in this paper, such approach might lead to a potential problem which is linked to the bandwidth selection for the continuous regressors due to the presence of the discrete regressors. Through the numerical study, we find that in many cases, the performance of the resulting nonparametric regression estimates may deteriorate if the discrete variables are smoothed in the way addressed so far in the literature, and that a fully separate estimation without any smoothing of the discrete variables may provide significantly better results. As a solution, we suggest a simple generalization of the popular approach proposed by Racine and Li [*Journal of Econometrics*, 2004] to address this problem and to provide estimates with more robust performance. We analyze the problem theoretically, develop the asymptotic theory for the new nonparametric smoothing method and study the finite sample behavior of the proposed generalized approach through extensive Monte-Carlo experiments as well present an empirical illustration.

Keywords: Discrete regressors, Nonparametric regression, Kernel smoothing, Cross-validation, Local linear smoothing

JEL Classification: C13, C14, C35

*Department of Mathematics, University of York, The United Kingdom; and Department of Econometrics and Business Statistics, Monash University, Australia; email: degui.li@york.ac.uk.

†Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium; email: leopold.simar@uclouvain.be.

‡School of Economics and Centre for Efficiency and Productivity Analysis, The University of Queensland, Australia; email: v.zelenyuk@uq.edu.au.

1 Introduction

Rapid advance of computing power and wide availability of large data sets encourage many researchers to substantially increase their attention to various nonparametric methods for estimating regression relationships. One of the most popular nonparametric methods appears to be the local polynomial least squares method, considered by Stone (1977), Cleveland (1979), Cleveland and Delvin (1988), Fan (1992, 1993), Fan and Gijbels (1992), Ruppert and Wand (1994) and popularized by Fan and Gijbels (1996). This method has received even greater appeal when it is substantially empowered by the seminal work of Racine and Li (2004), which suggests a neat way to deal with discrete regressors in the context of nonparametric regression.[1] This work has inspired many interesting applications in a wide range of areas, for example, by Stengos and Zacharias (2006), Maasoumi et al. (2007), Parmeter et al. (2007), Eren and Henderson (2008), Walls (2009), Hartarska et al. (2010), Henderson (2010), to mention just a few. In these and other works where Racine and Li (2004)'s approach is used, researchers are able to obtain new insights with much more confidence, as their approach is free from imposing any parametric form on the regression relationship, while using both continuous and discrete regressors without splitting the sample into sub-samples for each value of the discrete variables. In most of the empirical as well as theoretical studies and software codes using the Racine and Li (2004)'s approach that we are aware of, the way this approach is applied is in the "default" or simple form that we will describe below.

Indeed, it became somewhat common to smooth the discrete regressors in local polynomial least squares almost automatically, perceiving that one should obtain better results than if the nonparametric estimator is applied to each group separately (see Li and Racine, 2007; and the references therein). While this perception appears to hold in various cases and is supported by the simulations conducted by Racine and Li (2004) for the case of local constant fitting, in this article we illustrate that in some cases it may not be the case and smoothing the discrete variables can actually deteriorate substantially the resulting estimator of the regression function. In some situations even, a fully separate estimation for each group identified by the discrete (or categorical) variable may give more accurate results (e.g., in terms of Mean Squared Error, MSE) than the approach with smoothing over the discrete variable. In these cases, the reduction in variance, or the efficiency gain due to smoothing of the discrete regressors, can be well outweighed by a substantial bias introduced due to this smoothing. This may happen in both the small and relatively large samples, and so, for such cases, it may be preferable to make a fully separate estimation for each group.

[1] This paper extends the Aitchison and Aitken (1976)'s idea for smoothing discrete variables and establishes the asymptotic theory. Other basic references on smoothing discrete variables include Titterington (1980) and Wang and Van Ryzin (1981).

We will see below that the source of the problem comes from the bandwidth structure suggested by the basic or default method appearing in the existing literature which we are aware of (see, for example, some references listed above) and the various softwares implementing it. In this bandwidth selection procedure, a "simple" bandwidth scheme is proposed for the continuous variables, in the sense that the same bandwidth vector is taken across the various subgroups determined by the discrete variables, and then the bandwidths for the continuous and for the discrete variables are simultaneously determined. Hence, even if the resulting estimator of the bandwidth for a discrete variable takes the value of zero (i.e., no smoothing of this discrete variable with separate estimation by groups), the resulting bandwidths for the continuous variables are still restricted to be common to the various categories of the associated discrete variables. This may lead in some cases to "over-smoothing" in some groups and "under-smoothing" in the others.

To fix the ideas, suppose that we use local constant kernel method (Nadaraya-Watson) and that we have two groups of observations (determined by one discrete variable) and only one continuous variable. Suppose in addition that in one group, the variable is relevant (the derivative of the regression function with respect to this variable is not zero), and in the other it is not relevant. In the latter group, the optimal bandwidth would converge to infinity as shown by Hall et al. (2007), whereas in the first group, it may converge to zero. A fully separate analysis which indicates "non-smoothing" on the discrete variable would capture this feature whereas the "simple-smoothing" would miss such feature. The latter approach will provide a common bandwidth for the continuous variable that will under-smooth the regression in the group where the variable is irrelevant and over-smooth the regression in the group where it is relevant. This extreme case seems obvious but apparently it has been overlooked in practice. We can also imagine less extreme situations where the phenomenon would be similar: small influence of the continuous variable in one group and more complex structure in the other.

When the local polynomial smoothing is used, the same problem would also appear if in one group the local polynomial approximation is not far from the true regression, but the structure in the other is globally much more complex. For example, for the local linear case, the optimal bandwidth would converge to infinity if the true regression is linear, or converge to zero if the true regression function is highly-nonlinear (Li and Racine, 2004). Obviously, the problem can even be more severe when estimating the derivative of the nonparametric regression function. We will briefly illustrate this point in one example below, showing the consequences which the simple-smoothing method could lead to. To the best of our knowledge, this phenomenon has never been analyzed in the literature using smoothing techniques for discrete variables. It is one of the objectives of our paper to investigate the consequence of such issue.

In this paper, we also introduce a simple way to generalize and improve the commonly used smoothing methods and overcome the potential difficulties mentioned above. Specifically, in the bandwidth selection procedure, we allow different bandwidth parameters for the continuous variables in each category of the discrete variables. We call this method the "complete-smoothing" approach in contrast to the simple-smoothing approach used in most of the existing literature and to the non-smoothing approach where the groups are treated fully separately. This smoothing approach is, of course, at a cost of somewhat higher computational complexity, but we will see later that the gain in precision of the regression estimate can be substantial. We will limit the presentation to the case of categorical discrete variables where each value of the discrete variables determines a group.

More generally and beyond the extreme cases described above, whether the bias beats the variance or not, essentially depends on the degree of difference of curvatures of the regression relationship pertinent to each group identified by the discrete variable and to some extent also depends on other aspects of the Data Generating Process (DGP) such as the size of the noise. Thus, in general, *a priori* it is not clear whether it is better to smooth the discrete variables or to do a fully separate estimation (unless the latter is hardly reliable or impossible due to very small data in a given group) and, so far, there appears to be no formal rule of thumb for deciding on this dilemma. The complete-smoothing approach, that we suggest below, not only allows for smoothing the discrete variables, but also uses different bandwidths for each group for the continuous variables. Since this encompasses both the non-smoothing approach and the simple-smoothing technique as special cases, we can expect theoretically better performances of the complete-smoothing approach, which will be investigated later in the paper.

One of the main messages of our paper is that when using the simple-smoothing approach, an additional assumption on "similar degree of smoothness" with respect to continuous variables is implicitly imposed for different categories of the discrete variables, and one should recognize or acknowledge it explicitly. Moreover, we show that such additional assumption can distort the estimates substantially. While many examples can be used to illustrate the problem, we will deliberately take simple examples (univariate continuous and univariate discrete variable to avoid the potential curse of dimensionality problem), to vividly illustrate the point. We will also suggest a way to overcome this problem, whenever it is numerically doable, and point out the numerical difficulties involved there as well as open questions that still remain.

Because *a priori* it is not clear whether all categories of a discrete variable should have a common bandwidths for continuous variables or not, our finding is very important for practitioners, as it warns against an automatic use of smoothing over the discrete regressors in a simple way without considering that this might actually produce less accurate estimation

results. The title of our paper has a question mark and the question "To smooth or not to smooth?" remains open. So, we hope our paper will stimulate further research in this area to find some theoretically justified ways (e.g., statistical tests or rules of thumb) for deciding whether to smooth or not to smooth over the discrete regressors and whether or not to go with the more general method we suggest in this paper.

The rest of this paper is organized as follows. Section 2 recalls the basic methodologies of the local linear least-squares method[2]; Section 3 introduces the local linear complete-smoothing method and establishes the asymptotic theory; Section 4 illustrates how severe can be the problem by some simple visualized examples and by some more extensive Monte-Carlo experiments, and how the proposed complete-smoothing approach outperforms the other approaches; Section 5 illustrates the issue with a real data set and Section 6 concludes and summarizes our main findings.

2 Local Linear Simple-Smoothing

The point we aim to stress in this paper is a general phenomenon linked to nonparametric kernel-based regression, but we illustrate it on a method that appears to be the most popular in practice: the local linear least squares (LLLS) which is a particular case of local polynomial least squares (LPLS). We summarize here the basic idea of the method, and can refer to Fan and Gijbels (1996), Pagan and Ullah (1999), and Li and Racine (2007) for details. This method allows flexible form through approximating locally the true unknown regression. Assume that the dependent endogenous variables $Y_i \in \mathbb{R}$, $i = 1, \ldots, n$, are generated by the following regression model:

$$Y_i = m(Z_i) + \varepsilon_i, \tag{2.1}$$

where $Z_i = (Z_i^c, Z_i^d)$ with $Z_i^c \in \mathbb{R}^p$ being continuous and Z_i^d being a q-dimensional discrete vector. We next focus on the presentation for categorical unordered variables, but the same could be done for naturally ordered variables by using appropriate kernels, see Racine and Li (2004) for details. In addition, we make the standard assumptions on the errors, ε_i, that they are independent random variables with

$$\mathrm{E}(\varepsilon_i \mid Z_i) = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon_i \mid Z_i) < \infty \quad a.s.,$$

although the methodology we will discuss may also apply to more sophisticated setups. The flexibility of the model is related to the fact that the unknown regression function

[2]Our remarks and suggestions could obviously be applied to other nonparametric kernel-based estimation methods such as the local non-linear least squares, and local linear quasi-likelihood methods (see, for example, Gozalo and Linton, 2000; Frölich, 2006; Park et al., 2010).

$m(\cdot)$ is not specified. No particular assumptions are made on $m(\cdot)$ itself except for some smoothness properties on $m(\cdot, z^d)$. For the sake of simplicity, we assume that $m(\cdot, z^d)$ is twice continuously differentiable with respect to its first p continuous arguments. Finally, we need also some regularity condition on the smoothness of conditional density $f(z^c|z^d)$ with respect to the p continuous arguments.

The main idea of LPLS is to approximate $m(u, v)$ for all (u, v) in a neighborhood of a given point $z = (z^c, z^d)$ by a local polynomial function of degree r in the direction of z^c and then obtain the LPLS estimate by minimizing the resulting sum of the squared errors. A degree $r = 0$ would provide the local constant (Nadaraya-Watson type) estimator. As mentioned above, we will limit our presentation to the case of local linear approximation ($r = 1$). Extension to higher orders follows the same ideas but at a cost of more notational complexity. Consider the following local approximation:

$$m(u, v) \approx \alpha_{z^c,z^d} + \beta^{\tau}_{z^c,z^d}(u - z^c), \tag{2.2}$$

where $\alpha_{z^c,z^d} \in \mathbb{R}$ and $\beta_{z^c,z^d} \in \mathbb{R}^p$ are quantities to be estimated that, in general, vary with (z^c, z^d). To take only neighboring observations around (z^c, z^d), or to give more weights to them, when evaluating the least-squares criterion, the kernel approach is used. For the continuous variables we use a product kernel (but other multivariate kernels may also work), i.e.,

$$K_h(Z^c_i - z^c) = \prod_{j=1}^{p} \frac{1}{h_j} K_j\left(\frac{Z^c_{ij} - z^c_j}{h_j}\right), \tag{2.3}$$

where $h = (h_1, \ldots, h_p)$ is a vector of bandwidths, $K_j(\cdot)$ is a standard univariate kernel function such as the univariate standard Gaussian density, z^c_j is the j^{th} component of z^c and $Z^c_i = (Z^c_{i1}, \cdots, Z^c_{ip})^{\tau}$. For the discrete vectors, we use the discrete kernel introduced by Racine and Li (2004), i.e.,

$$\Lambda_\lambda(Z^d_i, z^d) = \prod_{\ell=1}^{q} \lambda_\ell^{I\left(Z^d_{i\ell} \neq z^d_\ell\right)}, \tag{2.4}$$

where $I(A)$ is the indicator function, with $I(A) = 1$ if A holds, and 0 otherwise, $\lambda_\ell \in [0, 1]$ are bandwidths for the discrete variables $\ell = 1, \ldots, q$, and $Z^d_i = (Z^d_{i1}, \cdots, Z^d_{iq})^{\tau}$. The LLLS or weighted least squares criterion at a given point (z^c, z^d) measuring the quality of the approximation is thus given by

$$C_n(\alpha_{z^c,z^d}, \beta_{z^c,z^d}; z^c, z^d) = \sum_{i=1}^{n} \left[Y_i - \alpha_{z^c,z^d} - \beta^{\tau}_{z^c,z^d}(Z^c_i - z^c)\right]^2 K_h(Z^c_i - z^c)\Lambda_\lambda(Z^d_i, z^d), \tag{2.5}$$

We note that if for a particular ℓ, we have $\lambda_\ell = 0$ (with the convention that $0^0 = 1$), then there is no smoothing of this ℓ^{th} discrete variable; i.e., the evaluation in (2.5) is done separately for each subsample determined by this discrete variable, with a common h. At the

other limit, if $\lambda_\ell = 1$, we do not take into account the discrete variable $Z_{i\ell}^d$ in the analysis, i.e., all the sample points have weight in (2.5) which is independent from the value of $Z_{i\ell}^d$. Let $\widehat{\alpha}_{z^c,z^d}$ and $\widehat{\beta}_{z^c,z^d}$ minimize the criterion $C_n(\cdot,\cdot;z)$ with $z=(z^c,z^d)$, then the proposed estimated value of the regression function at the point (z^c,z^d), denoted by $\widehat{m}(z^c,z^d)$, is given by $\widehat{\alpha}_{z^c,z^d}$ whereas $\widehat{\beta}_{z^c,z^d}$ gives the estimated value of the first partial derivative of $m(\cdot)$ with respect to the continuous variables z^c evaluated at (z^c,z^d). As a common bandwidth is used to smooth over the continuous variables for different subgroups, we call such method as local linear simple-smoothing.

The bandwidth selection is a very important issue in nonparametric kernel-based estimation. The selection of appropriate bandwidths (h,λ) in (2.5) can be done by the cross-validation (CV) method, although many other approaches can be adopted such as the corrected AIC method. When adopting the CV approach, the values $\hat{h}$ and $\hat{\lambda}$ are chosen to minimize

$$\mathrm{CV}(h,\lambda)=\sum_{i=1}^{n}\left[Y_i-\widehat{m}_{(-i)}(Z_i^c,Z_i^d\big|h,\lambda)\right]^2 M(Z_i^c,Z_i^d), \tag{2.6}$$

where $M(Z_i^c,Z_i^d)$ is a weight function trimming out boundary observations and $\widehat{m}_{(-i)}(Z_i^c,Z_i^d)$ is the leave-one-out kernel estimator of $m(Z_i^c,Z_i^d)$ with bandwidths h and λ, i.e., estimated by minimizing (2.5), but leaving the i^{th} observation out of the sample. The properties of the resulting estimator $\widehat{m}(z^c,z^d)$ by using the CV bandwidth selection method have been described in the seminal paper by Racine and Li (2004) for the local constant case ($r=0$), and in Li and Racine (2004) for the local linear approximations. It is important to notice that, as common in the nonparametric smoothing approaches, these results assume that $h_j \to 0$ for all $j=1,\ldots,p$, and $nh_1\ldots h_p \to \infty$ as $n\to\infty$. The argument generally admitted in the existing literature is that it is better to smooth the discrete variable in (2.5), because the separate analysis on each separate subsample defined by the categorial variables Z^d would correspond to the particular case when all $\lambda_\ell = 0$, $\ell=1,\ldots,q$. However, we indicate in the next section that it might not be the case in some situations.

3 Local Linear Complete-Smoothing

To simplify the argument, let us suppose that we have univariate discrete variable defining two groups of observations (the argument would be the same when considering all the subgroups defined by the multivariate categorical vector Z^d). The DGP in the two groups may have different characteristics (shape of the regression function, curvature, or size of the noise). Extreme cases of such differences have been shortly described in the introductory section. Unless the sample size within one of the groups is very small, the CV bandwidth

selection procedure described above would provide small value of tuning parameter λ for relevant discrete regressor, resulting in the frequency method in the limiting case where $\lambda = 0$, and a common h for different groups. Furthermore, we also suppose that the continuous variable Z^c is also univariate. Extension to the case of multivariate continuous and discrete variables is straightforward. However, such extension may not be so useful because of the curse of dimensionality problem.

As pointed out above for the extreme cases, the simple-smoothing based CV method might be inappropriate in many situations where some important characteristics of the DGP are quite different between the two subgroups. In such instances, having common bandwidth h for different groups, may lead to under-smoothing (over continuous variables) for one group while over-smoothing for the other. While this may not matter asymptotically as long as the common bandwidth has the proper order, it happens to matter in finite sample case (small and even relatively large ones), sometimes substantially, as illustrated in our examples in the next section. To address this issue, it might be better to do fully separate estimation within each subgroup, allowing the bandwidth over continuous variable to vary across different groups without smoothing for the discrete variable. Although, as pointed out by Li and Racine (2004), this may increase the variance, the bias could be lowered substantially, which would lead to the estimation with smaller MSE than the simple-smoothing method. We will show in the next section, through the Monte-Carlo simulations, that the loss in accuracy of the default simple-smoothing estimation method may be dramatic. However, it is well-known that if the sample size in one group is too small, one cannot hope to get sensible results with the separate local linear estimation. We next give a solution to address this problem.

A natural way is to allow different bandwidths for the continuous variable in the different subgroups which address the problem of overcoming over- and under-smoothing in subgroups, and allow smoothing over the discrete variable, as in Racine and Li (2004), to circumvent the problem that the number of observations in one subgroup might be too small. We call such method as the complete-smoothing method. Formally, in the case of two subgroups defined by one discrete variable, the equation (2.5) defining the estimator could be replaced by

$$C_n(\alpha_{z^c,z^d}, \beta_{z^c,z^d}; z^c, z^d) = \sum_{i=1}^{n} \Big[Y_i - \alpha_{z^c,z^d} - \beta^{\tau}_{z^c,z^d}(Z_i^c - z^c)\Big]^2 \Lambda_\lambda(Z_i^d, z^d)$$
$$\times \Big[K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\} + K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\}\Big], \qquad (3.1)$$

where $z^d(k)$, $k = 1, 2$, are the two possible values of z^d. For simplicity, we use the same kernel function in the two subgroups, but allow potentially different bandwidths $h(1)$ and $h(2)$ for these subgroups. Let $\widetilde{m}(z^c, z^d)$ be the local linear estimated value of $m(z^c, z^d)$

with complete-smoothing which minimize $C_n(\alpha_{z^c,z^d},\beta_{z^c,z^d};z^c,z^d)$ with respect to α_{z^c,z^d} and β_{z^c,z^d}. It is clear that the general formulation we propose in (3.1) encompasses both the fully separate analysis by groups ($\lambda=0$), and the simple-smoothing approach ($h(1)=h(2)$).

Let $\mu_j=\int u^jK(u)du$ and $\nu_j=\int u^jK^2(u)du$. Define $\sigma_m^2(z^c,z^d)=\frac{\nu_0\sigma^2(z^c,z^d)}{f(z^c|z^d)P(Z^d=z^d)}$ and

$$\begin{aligned} b_{m,1}(z^c) &= \frac{1}{2}\mu_2 m''(z^c,z^d(1))h^2(1)+\lambda(1-p_1)\big[m(z^c,z^d(2))-m(z^c,z^d(1))\big]/p_1,\\ b_{m,2}(z^c) &= \frac{1}{2}\mu_2 m''(z^c,z^d(2))h^2(2)+\lambda p_1\big[m(z^c,z^d(1))-m(z^c,z^d(2))\big]/(1-p_1), \end{aligned}$$

where $\sigma^2(\cdot,\cdot)$ and $f(\cdot|\cdot)$ are defined in Assumption 3 in Appendix A and $0<p_1=P(Z^d=z^d(1))<1$. We next give the asymptotic distribution theory for the local linear complete smoothing estimator $\widetilde{m}(z^c,z^d)$.

THEOREM 3.1. *Suppose that Assumptions 1–4 in Appendix A are satisfied. If $z^d=z^d(1)$, we have*

$$\sqrt{nh(1)}\Big[\widetilde{m}(z^c,z^d(1))-m(z^c,z^d(1))-b_{m,1}(z^c)\Big]\xrightarrow{d}\mathrm{N}\big[0,\sigma_m^2(z^c,z^d(1))\big]. \tag{3.2}$$

If $z^d=z^d(2)$, we have

$$\sqrt{nh(2)}\Big[\widetilde{m}(z^c,z^d(2))-m(z^c,z^d(2))-b_{m,2}(z^c)\Big]\xrightarrow{d}\mathrm{N}\big[0,\sigma_m^2(z^c,z^d(2))\big]. \tag{3.3}$$

We next derive the optimal bandwidths for the local linear complete smoothing estimator. Let $\widetilde{m}_{(-i)}(Z_i^c,Z_i^d,h(1),h(2),\lambda)$ be the leave-one-out local linear complete-smoothing estimated value of $m(Z_i^c,Z_i^d)$ with tuning parameters $h(1)$, $h(2)$ and λ. Then, similarly to the CV method introduced in Section 2, we define

$$\overline{\mathrm{CV}}(h(1),h(2),\lambda)=\sum_{i=1}^n\big[Y_i-\widetilde{m}_{(-i)}(Z_i^c,Z_i^d\big|h(1),h(2),\lambda)\big]^2M(Z_i^c,Z_i^d). \tag{3.4}$$

The optimal bandwidths $[\hat{h}(1),\hat{h}(2),\hat{\lambda}]$ are the values which minimize $\overline{\mathrm{CV}}(h(1),h(2),\lambda)$. Define

$$\begin{aligned} \psi_1(h(1),\lambda) &= p_1\int\Big\{b_*(z^c,z^d(1))h^2(1)+\frac{\lambda(1-p_1)\big[m(z^c,z^d(2))-m(z^c,z^d(1))\big]}{p_1}\Big\}^2\\ &\quad\times M(z^c,z^d(1))f(z^c|z^d(1))dz^c,\\ \psi_2(h(2),\lambda) &= (1-p_1)\int\Big\{b_*(z^c,z^d(2))h^2(2)+\frac{\lambda p_1\big[m(z^c,z^d(1))-m(z^c,z^d(2))\big]}{1-p_1}\Big\}^2\\ &\quad\times M(z^c,z^d(2))f(z^c|z^d(2))dz^c, \end{aligned}$$

where $b_*(z^c, z^d) = \frac{1}{2}\mu_2 m''(z^c, z^d)$. Define

$$\begin{aligned}\chi(h(1)) &= \frac{\nu_0}{h(1)}\int \sigma^2(z^c, z^d(1))M(z^c, z^d(1))dz,\\ \chi(h(2)) &= \frac{\nu_0}{h(2)}\int \sigma^2(z^c, z^d(2))M(z^c, z^d(2))dz.\end{aligned}$$

We next give the asymptotic expansion of $\overline{\mathrm{CV}}(h(1), h(2), \lambda)$, which is critical to derive the asymptotic property of the optimal bandwidths $[\hat{h}(1), \hat{h}(2), \hat{\lambda}]$.

THEOREM 3.2. *Suppose that the conditions in Theorem 3.1 and Assumption 4′ are satisfied. Then, we have*

$$\overline{\mathrm{CV}}(h(1), h(2), \lambda) = \overline{\mathrm{CV}}_1 + \Phi(h(1), h(2), \lambda) + s.o., \tag{3.5}$$

where $\overline{\mathrm{CV}}_1 := \sum_{i=1}^n \varepsilon_i^2 M(Z_i^c, Z_i^d)$ *is independent of the tuning parameters,*

$$\Phi(h(1), h(2), \lambda) = n\big[\psi_1(h(1), \lambda) + \psi_2(h(2), \lambda)\big] + \chi(h(1)) + \chi(h(2)),$$

and s.o. represents some terms with smaller asymptotic order.

Define

$$\overline{\mathrm{MSE}}(h(1), h(2), \lambda) = \sum_{i=1}^n \big[m(Z_i^c, Z_i^d) - \widetilde{m}_{(-i)}(Z_i^c, Z_i^d\big|h(1), h(2), \lambda)\big]^2 M(Z_i^c, Z_i^d). \tag{3.6}$$

In the proof of Theorem 3.2 in Appendix B, we show that

$$\overline{\mathrm{MSE}}(h(1), h(2), \lambda) = \Phi(h(1), h(2), \lambda) + s.o. \tag{3.7}$$

Letting $[h_*(1), h_*(2), \lambda_*]$ be the minimizer to $\overline{\mathrm{MSE}}(h(1), h(2), \lambda)$, by Theorem 3.2 and standard argument (such as the proof of Theorem 3.1 in Li and Racine, 2004), we have

$$\frac{\hat{h}(1) - h_*(1)}{h_*(1)} = o_P(1), \quad \frac{\hat{h}(2) - h_*(2)}{h_*(2)} = o_P(1), \quad \frac{\hat{\lambda} - \lambda_*}{\lambda_*} = o_P(1). \tag{3.8}$$

We next compare the measurements of the MSEs between local linear complete-smoothing and local linear simple-smoothing. Let

$$\mathrm{MSE}(h, \lambda) = \sum_{i=1}^n \big[m(Z_i^c, Z_i^d) - \widehat{m}_{(-i)}(Z_i^c, Z_i^d\big|h, \lambda)\big]^2 M(Z_i^c, Z_i^d). \tag{3.9}$$

and $[h_0, \lambda_0]$ be the minimizer to $\mathrm{MSE}(h, \lambda)$. Analogously, we can also show that

$$\frac{\hat{h} - h_0}{h_0} = o_P(1), \quad \frac{\hat{\lambda} - \lambda_0}{\lambda_0} = o_P(1). \tag{3.10}$$

Noting that

$$\overline{\mathrm{MSE}}(h(1), h(2), \lambda) = \mathrm{MSE}(h, \lambda) + s.o.$$

for the case of $h(1) = h(2) = h$ and $\lambda = o(h^2)$, we may show that

$$\min_{h(1),h(2),\lambda} \overline{\mathrm{MSE}}(h(1), h(2), \lambda) \leq \min_{h,\lambda} \overline{\mathrm{MSE}}(h, \lambda). \tag{3.11}$$

Using (3.8), (3.10) and (3.11), we can show that

$$\overline{\mathrm{MSE}}(\hat{h}(1), \hat{h}(2), \hat{\lambda}) \leq \mathrm{MSE}(\hat{h}, \hat{\lambda}),$$

which indicates that in the large sample case, the MSE by using the optimal bandwidths chosen by our method is smaller than that by using the default optimal bandwidths.

In the next section, we will investigate the finite sample properties of the different approaches. This will further confirm the expected theoretical performance of our approach over the traditional smoothing approach and over the fully separate approach. In particular, we find that the gain of precision may be substantial in practice.

4 Simulation Studies

In this section, we first present some simple examples that allow us to vividly illustrate the issue raised in the previous sections. We provide some "typical" pictures resulting from these simulated samples generated according to the scenario described below. Of course we cannot conclude general statements with one simulated sample, but the idea is to provide visualization of the problem. We then confirm what we see by a more detailed Monte-Carlo experiment. Throughout the simulation, we considered the following regression relationship

$$Y_i = a_1 + a_2 Z_i^d + b_1 Z_i^c + b_2 Z_i^d Z_i^c + b_3 (Z_i^c)^2 + b_4 Z_i^d (Z_i^c)^2 + b_5 Z_i^d \sin(\pi Z_i^c) + \varepsilon_i. \tag{4.1}$$

We have also tried some other regression relationships in simulation, and obtained similar conclusions. Varying the choice of the parameters in model (4.1) would lead to various examples explored in this section.

Example 1: linear versus periodic regression

Let $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 0, b_5 = 2$, with $\varepsilon_i \sim \mathrm{N}(0, \sigma_{\varepsilon,i})$, where $\sigma_{\varepsilon,i} = 2 - Z_i^d$. Here, for each simulation, the $Z_i^c \sim \mathrm{U}(-2, 2)$ for the continuous variable Z^c and the discrete variable Z^d was set randomly at 1 if $W > 0.25$ and set at 0 if $W \leq 0.25$, where $W \sim \mathrm{U}(0, 1)$. So, we randomly obtained about 75% of observations for group 1 (with $Z_i^d = 1$) and about 25% for group 2 (with $Z_i^d = 0$).[3]

In this example, for group 2 ($z^d = 0$) we have a linear model with more noisy data and smaller sample size, and for group 1 ($z^d = 1$) we have a linear model (with different intercept

[3]Most conclusions remain the same for other compositions (e.g., 50% versus 50%).

and slope) plus a cyclical component. In Figure 1, we present typical results of the estimation by using the three approaches: the fully separate estimation of the two subsamples (i.e., non-smoothing over the discrete variable, Approach 1), the simple-smoothing (i.e., smoothing the discrete variable with common bandwidth for the continuous regressor, Approach 2), and the complete-smoothing (i.e., smoothing the discrete variable and keeping potentially different bandwidths for the continuous variable, Approach 3). Only two cases of sample size are provided in Figure 1: $n = 100$ and $n = 400$ (similar pictures have been obtained for other sizes).

From the left panels ($n = 100$), we can see that the simple-smoothing (Approach 2) suffers from a serious drawback in this scenario and that the non-smoothing estimation (Approach 1) gives much better results both for $n = 100$ and $n = 400$. The complete-smoothing of Approach 3, encompassing the 2 preceding ones, does as well as the fully separate analysis for these samples. The fully separate estimation substantially outperforms the estimation with simple-smoothing over the discrete regressor, as the latter approach under-smoothes for group 2 and slightly over-smoothes for group 1. Note that for this example, the under-smoothing is more pronounced because group 1 dominates in the pooled sample by its larger size and so the common bandwidth selected in the CV optimization for (h, λ) is relatively close to what is optimal for group 1 in the separate estimation, while for the group 2, the true optimal bandwidth must in fact go to infinity to attain the correctly specified parametric model. The complete-smoothing (Approach 3) avoids such problem by allowing different bandwidths for the continuous regressor Z^c in the two groups. One may argue that this problem is caused the relatively small sample size in Figure 1. However, when looking at the left panel of Figure 4, one can see that when $n = 1000$, the performance of Approach 2 is still poor for group 2 due to the persistent under-smoothing. In contrast, with such large samples, allowing different bandwidths for the continuous variable in the two groups (Approaches 1 and 3) gives substantially improved regression fitting.

The numerical results of the Monte-Carlo experiment summarized in Table 5 further confirm the above analysis from Figure 1. With $n = 50, 100, 200, 400$, we conducted $M = 500$ Monte-Carlo replications for each case. The table provides the mean of the Approximate Mean Squared Error (AMSE) for each sample, averaged over 500 Monte-Carlo replications, i.e., $\overline{\text{AMSE}} = (1/M)\sum_{g=1}^{M} \text{AMSE}_g$, where the AMSE for each replication g ($g = 1, \ldots, M$) is defined as

$$\text{AMSE}_g = \frac{1}{n}\sum_{i=1}^{n}\left[m(Z^c_{i,g}, Z^d_{i,g}) - \widehat{m}(Z^c_{i,g}, Z^d_{i,g})\right]^2.$$

Table 5 also gives the estimated standard deviation of $\overline{\text{AMSE}}$ defined as

$$std_{MC} = \sqrt{\frac{1}{M(M-1)}\sum_{g=1}^{M}\left(\text{AMSE}_g - \overline{\text{AMSE}}\right)^2},$$

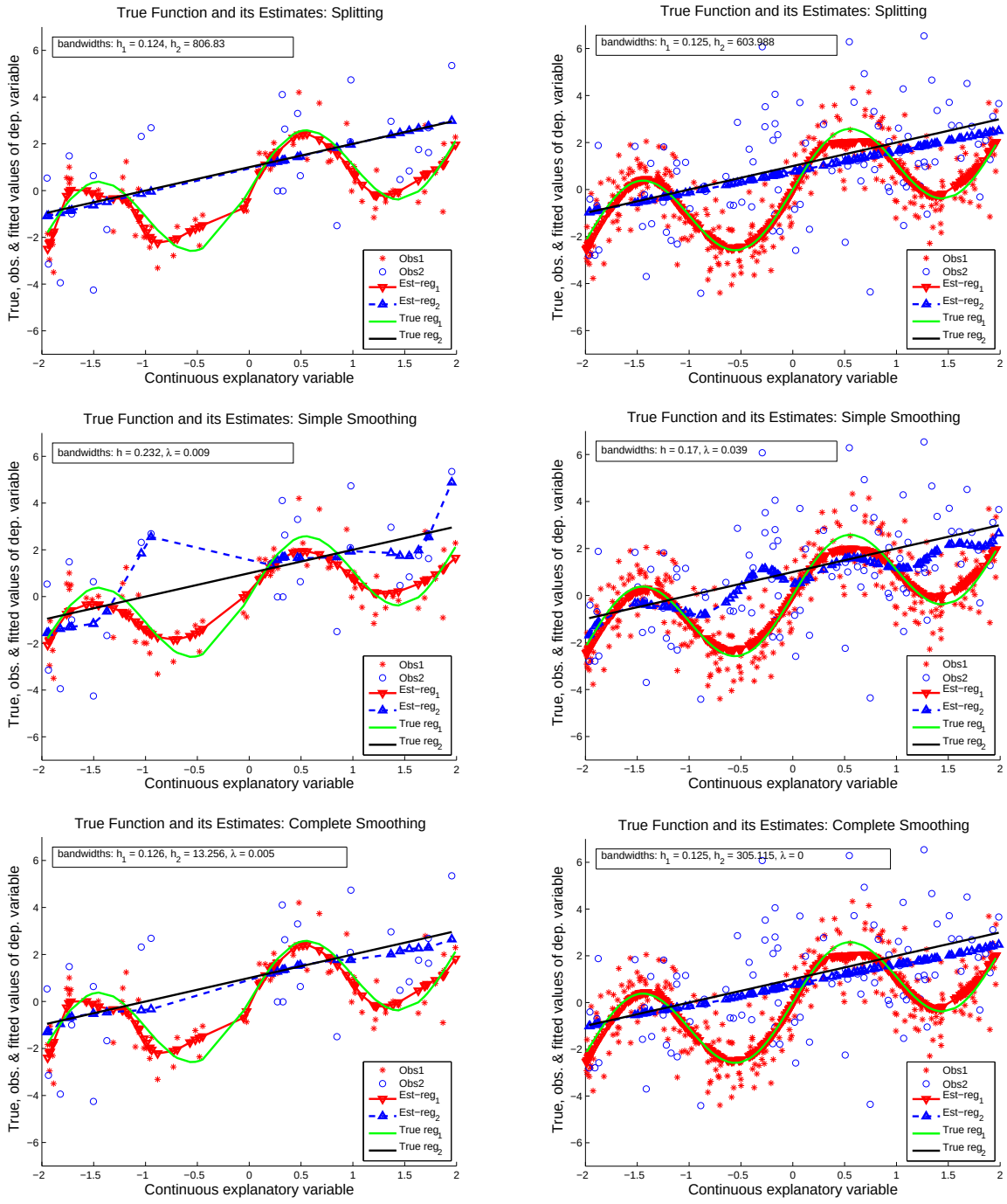

Figure 1: *Example 1: Left panel, $n = 100$ and right panel, $n = 400$. From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple-smoothing), Approach 3 (local linear complete-smoothing).*

This std_{MC} can be used to check if the differences observed in the table for the $\overline{\text{AMSE}}$ are

significant. To save space, "Split", "Simpl" and "Compl" represent Approach 1, Approach 2 and Approach 3, respectively.

Note that all the figures appearing in Table 5 vary as expected when n increases. It is worth noting that for all the simulations in this scenario, the CV method yielded $\hat{\lambda}$ that is very close to zero for both Approach 2 and Approach 3, and that Approach 3 gave systematically less weight in the smoothing of the discrete variable than Approach 2 (in terms of medians over the 500 replications). In Table 5, we can also see that the $\overline{\text{AMSE}}$ of Approach 1 is always smaller than that of Approach 2 (often about twice smaller), and that this does not vanish when the sample size increases. Note that the difference in $\overline{\text{AMSE}}$ is much larger for the smaller group, which has been explained above based on the findings of Figure 1. Meanwhile, from the medians of the bandwidth selected by the different approaches, we see clearly that Approach 2 under-smoothes continuous variable for Group 2. Furthermore, we can see that the complete-smoothing approach is, as expected and explained above, very robust here since it gives almost the same results as Approach 1. Note also that, as a consequence of the theory provided by Racine and Li (2004) and Li and Racine (2004), in all the cases, the AMSE reduces as n increases and that the optimal bandwidths (except $\hat{h}(2)$ when computed separately) go to zero as n goes to infinity. On the other hand, $\hat{h}(1)$ and $\hat{h}(2)$ chosen by the CV method using Approach 3 are similar to those obtained using Approach 1, and the optimal $\hat{\lambda}$ by using Approach 3 is smaller than that using Approach 2, indicating that Approach 2 suggests more similarities between groups than Approach 3.

Example 2: linear versus quadratic regression

In this example, we took much more similar regression lines for the two groups, compared with the previous example. In model (4.1), we selected the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 1, b_5 = 0$ with other assumptions remaining the same as those in Example 1. So in this example, for group 1 ($Z_i^d = 1$) we have replaced the periodic term in Example 1 by a quadratic term, which implies certain curvature. The scenario for group 2 ($Z_i^d = 0$) remains the same as that in the first example.

Figure 2 shows typical examples with the resulting fits of the three approaches, with $n = 100$ and $n = 400$. The plots tell us the same story as in Example 1, although the difference in curvature is not so radical. This is also confirmed by the numerical results of 500 Monte-Carlo experiments displayed in Table **??**. To summarize and to save space, we just note that most of the comments coming from Example 1 can be replicated here: Approaches 1 and 3 give better results than Approach 2, with AMSE often being about two or even three times smaller. The medians of CV-estimated values of λ are substantially larger for Approach 2 than that for Approach 3, indicating that Approach 2 assigns more similarity to the different groups than Approach 3 does.

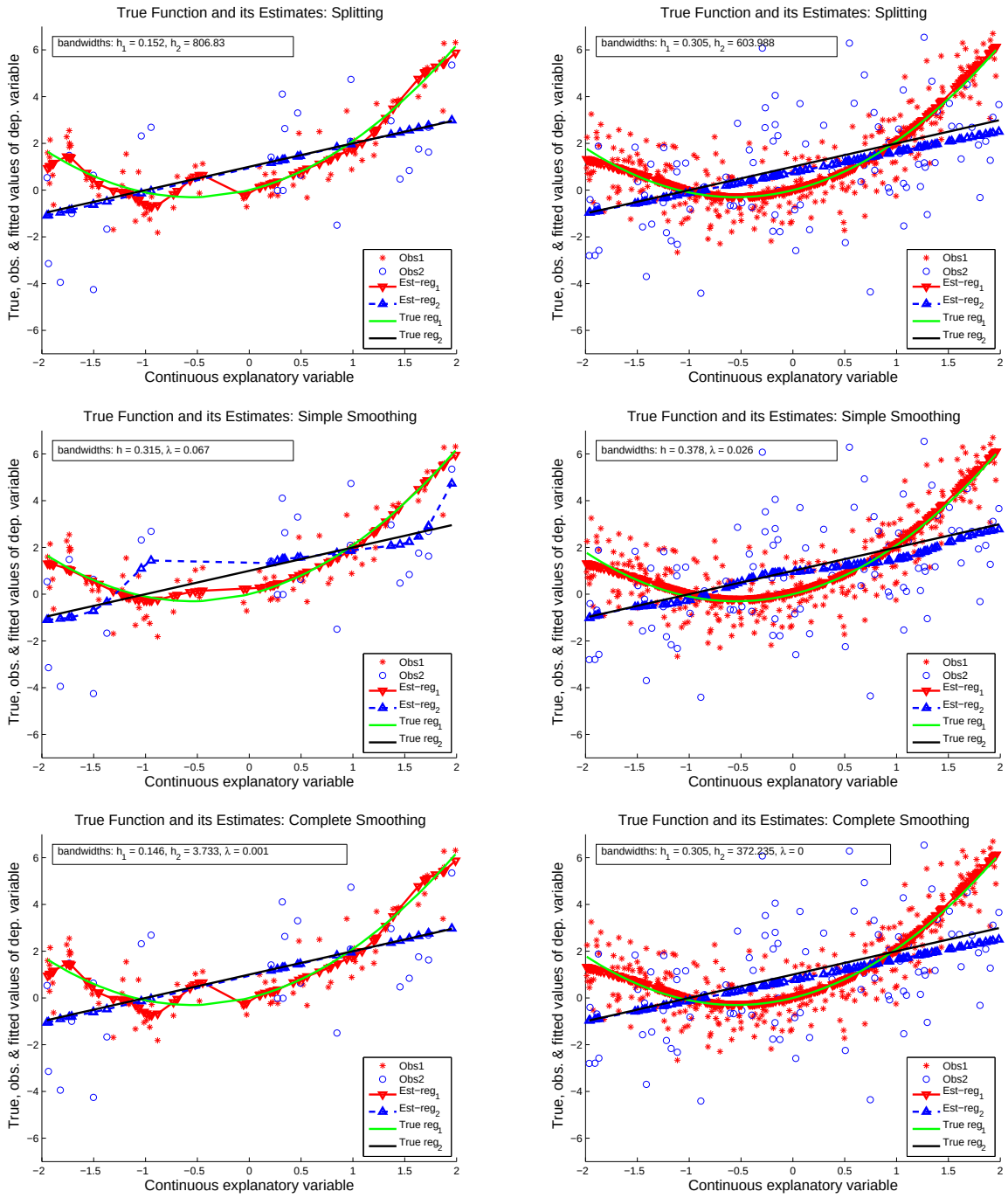

Figure 2: *Example 2: Left panel, $n = 100$ and right panel, $n = 400$. From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple-smoothing), Approach 3 (local linear complete-smoothing).*

Example 3: linear versus very similar quadratic regression

We next describe another scenario where the quadratic regression is quite similar to the linear one: the size of the intercept is changed from -1 to -0.5 and the coefficient of quadratic term is decreased from 1 to 0.25. So, in model (4.1), we selected the values $a_1 = 1, a_2 = -0.5, b_1 = 1, b_2 = 0.1, b_3 = 0, b_4 = 0.25, b_5 = 0$ with all the other scenario remaining the same. We do not present the pictures because the two regression lines are so close that we do not learn too much by looking into a particular sample realization. However, the Monte-Carlo numerical results presented in Table **??** are not without interest.

As expected, the simple-smoothing technique (Approach 2) generally behaves significantly better than Approach 1, especially for the case of $n = 50$. This is coherent with simulation results from Racine and Li (2004). In some cases it is also significantly better than Approach 3, but the difference is usually negligible and is never in the magnitudes observed in previous example. Approach 2 here outperforms the non-smoothing case (Approach 1), mostly due to the gain of precision in estimating the regression relationship for the group 2. The gain of variance is larger than the loss due to bias. However, the dominance of Approach 2 over Approach 1 vanishes very quickly as n increases (already negligible differences for the case of $n = 200$).

Example 4: quadratic versus quadratic regression

In this example, both regression relationships are quadratic, which indicates that the theoretical optimal values of the bandwidths in both groups should converge to zero as the sample size increases. We keeped the difference in the intercept between the two regressions as in Examples 1 and 2, and a slight difference of the linear component, but one regression has a quadratic component which is two times larger than the other. Specifically in equation (4.1), we took the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0.15, b_4 = 0.15, b_5 = 0$, with all the other scenario remaining the same as in Example 1. As in Example 3, it is expected to have a better behavior of the simple-smoothing technique (Approach 2) than the non-smoothing technique (Approach 1), because the curvatures of the two regressions are fairly similar.

Figure 3 shows the resulting fits of the three approaches, with $n = 100$ and $n = 400$. We see indeed that Approach 2 behaves slightly better than the other two approaches, although group 2 seems again to be slightly under-smoothed because of the same reason explained above in Examples 1 and 2. From the right panel of Figure 4, we find that even for the relatively large sample case ($n = 1000$), Approach 2 does not provide estimators so close to the true regressions, as Approaches 1 and 3 do. Table **??** indicates that Approach 2 does mostly as well as Approach 1 and sometimes better ($n = 50$), whenever the total $\overline{\text{AMSE}}$ is considered, which is at a cost of balancing a better behavior for group 2 and a worse behavior for group 1. On the other hand, Approach 3 gives almost the same results as Approach 1, and slightly better performances than Approach 2 for $n = 400$.

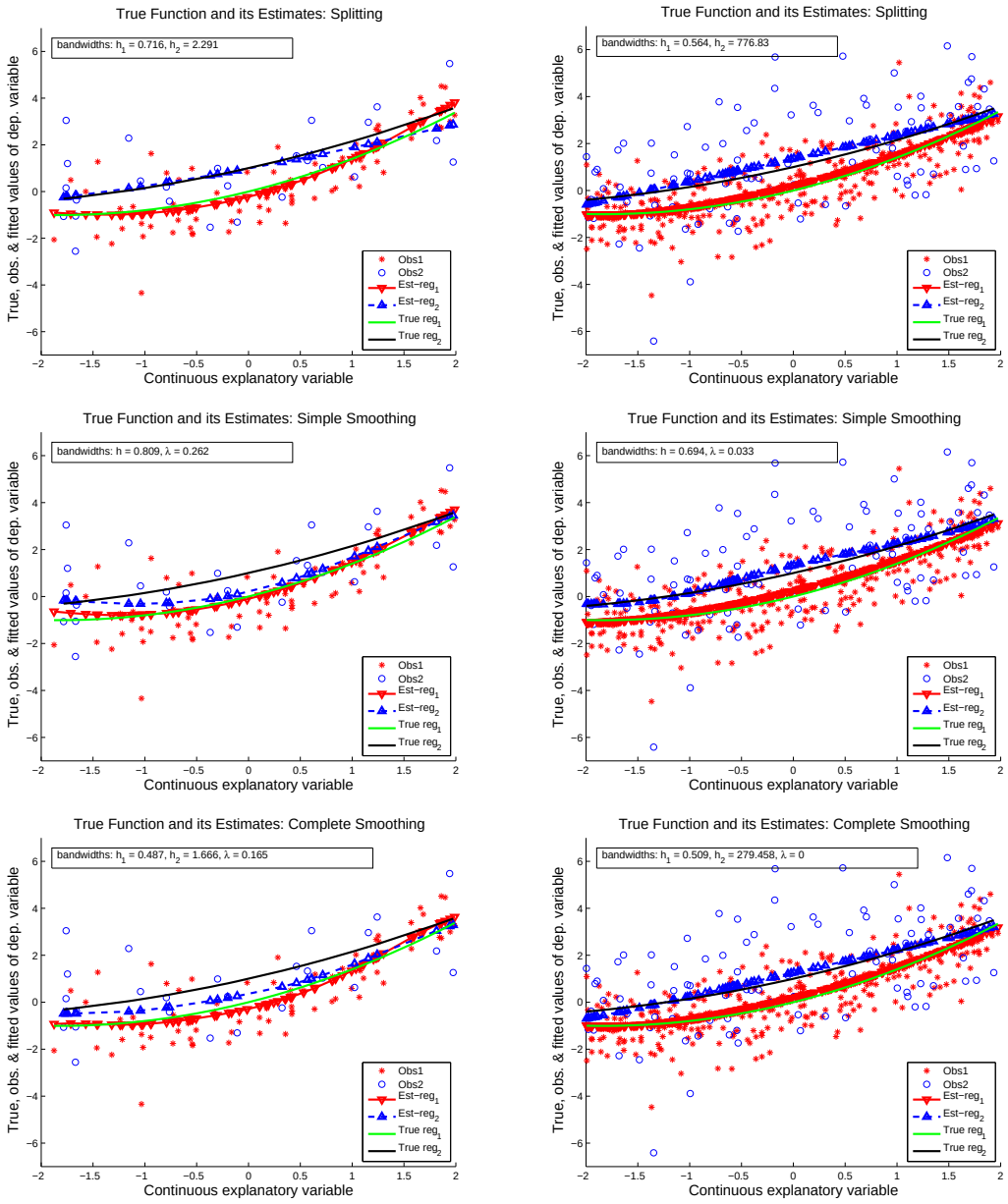

Figure 3: *Example 4: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple-smoothing), Approach 3 (local linear complete-smoothing).*

Example 5: quadratic versus periodic regression

We next consider a quadratic versus a periodic regression in the two groups, which is similar to Example 1 except that now the linear model is wrong in both groups. We selected in equation (4.1) the values $a_1 = 1, a_2 = -1, b_1 = 1, b_2 = 0.1, b_3 = 0.25, b_4 = 0, b_5 = 2$, with all the other assumptions remaining the same as in Example 1. The simulation results are

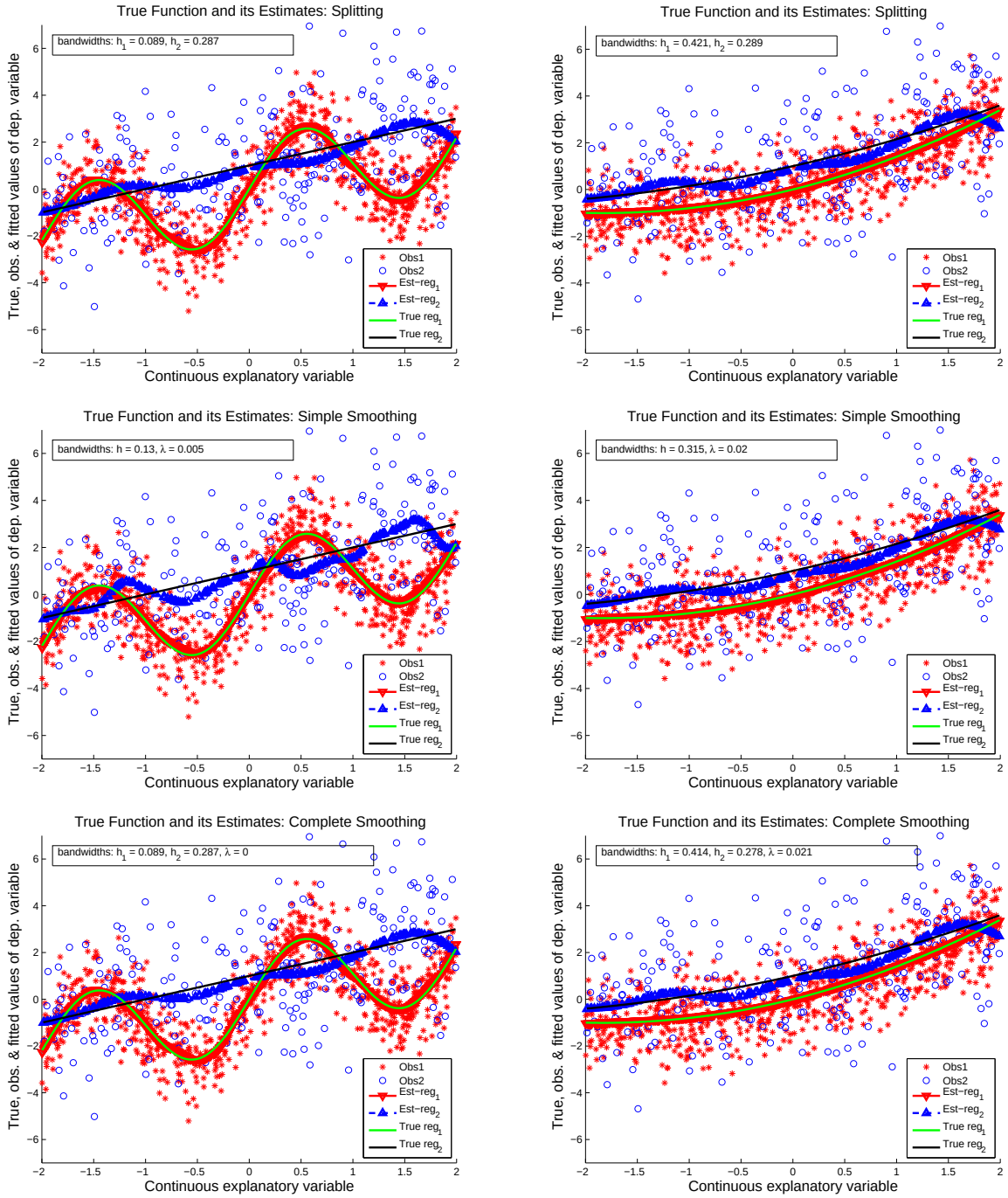

Figure 4: *Left panel, Example 1 with* $n = 1000$ *and right panel, Example 4 with* $n = 1000$. *From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple-smoothing), Approach 3 (local linear complete-smoothing).*

given in Figure 5 and Table **??**. We find that in this case the difference between Approach 2 and the other two approaches is still present. From Table **??**, we confirm the general

comments given for Example 1: Approach 1 has the better performance than Approach 2, and Approach 3 is the more robust way in this scenario (with significantly better performances). In particular, note that even though the difference in curvatures between the two groups is not so extreme as in Example 1, the difference of performances is still substantial: the overall AMSE for Approach 2 is about 1.5 times larger than those for Approaches 1 and 3 when $n = 50$, and about twice larger for $n = 100$ and larger samples.

Consequences on the estimation of derivatives

The estimation of derivatives also has the phenomenon similar to what we just described. To save space, we only illustrate this for the case of the scenario described in Example 1. Figure 6 displays one typical sample and the resulting estimates of the first partial derivatives using the three approaches. This figure shows that the estimation of derivatives can be even more severely flawed by using the simple-smoothing approach. Indeed, as one can clearly see from Figure 6, with simple-smoothing technique, one obtained radically varying estimated curves of the derivatives for the group where their true values are constant. The problem sustains whether the total sample size is 100 or 400 (or more). This means that research conclusions, policy implications and, consequently, the real policy decisions based on such estimates can be misleading. Note that for the same example, the complete-smoothing approach produced much better results which are very close to the true values. We also did a more complete Monte-Carlo experiment that confirmed largely these expected results, but we omit them from the paper to save space.

5 Empirical Application

In this section, we make an illustration of the phenomenon discussed above in the context of a real data set from the study by Kumar and Russell (2002), about patterns of convergence or divergence in economic growth in the World.[4] We chose this data and the context because the topic of economic growth has remained interesting for a wide audience for centuries.

This data set consists of observations in 57 countries, containing the variables such as GDP, labour and capital of each country in 1965 and in 1990, and was originally extracted from the Penn World Tables. We will use this data to estimate regression relationship between the growth in GDP per capita of countries between 1965 and 1990 (response variable) and the initial levels of GDP per capita of these countries (continuous explanatory variable). Such a regression and many of its variations are often performed in empirical economic

[4]This data set (or its extended version) was also used in many other applications, see, for example, Henderson and Russell (2005), Simar and Zelenyuk (2006), Henderson and Zelenyuk (2007) and Badunenko et al. (2008).

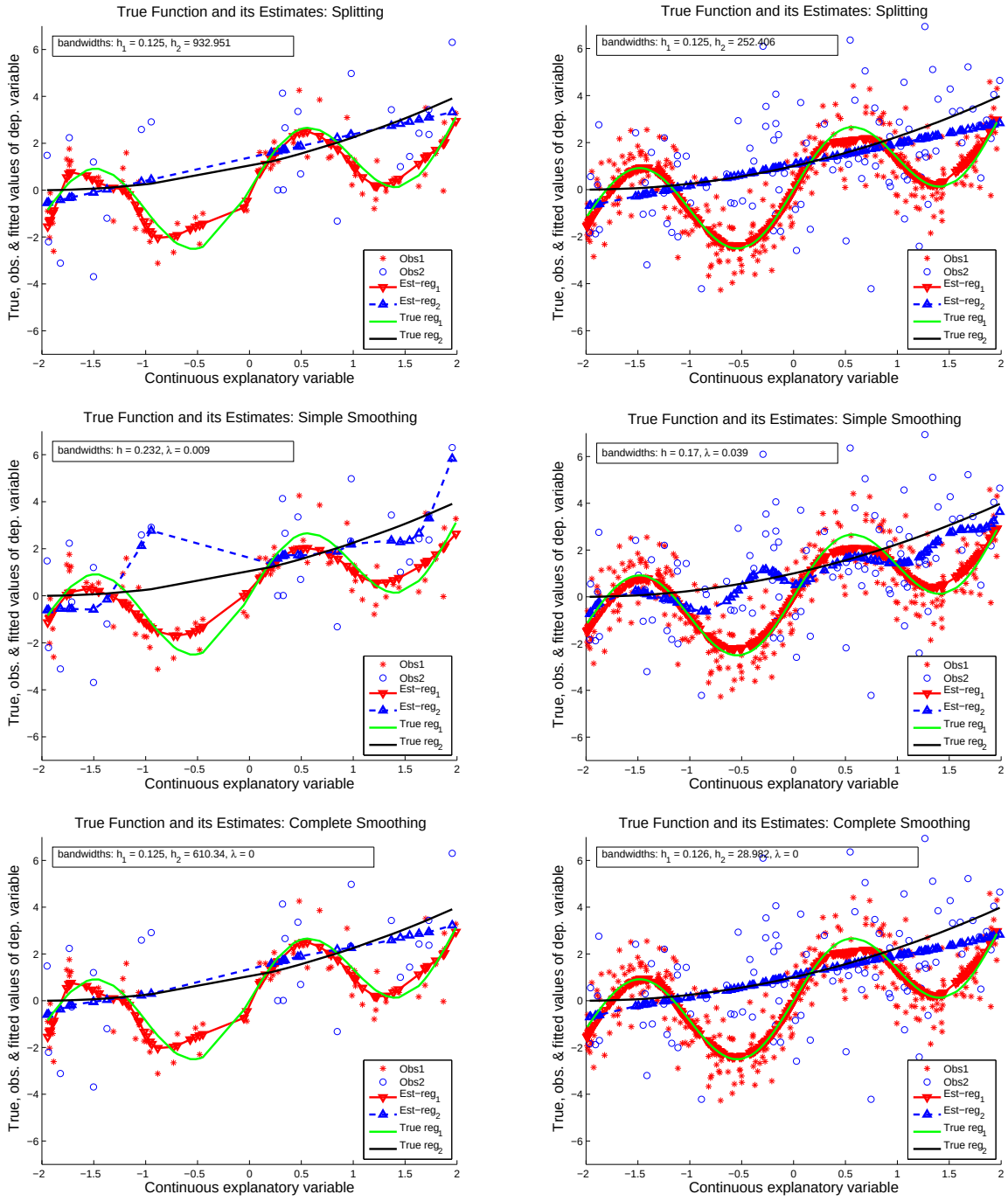

Figure 5: *Example 5: Left panel, $n = 100$, and right panel, $n = 400$. From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple-smoothing), Approach 3 (local linear complete-smoothing).*

growth studies on convergence.

The interest of such studies often lies on that the growth rates of poorer countries are,

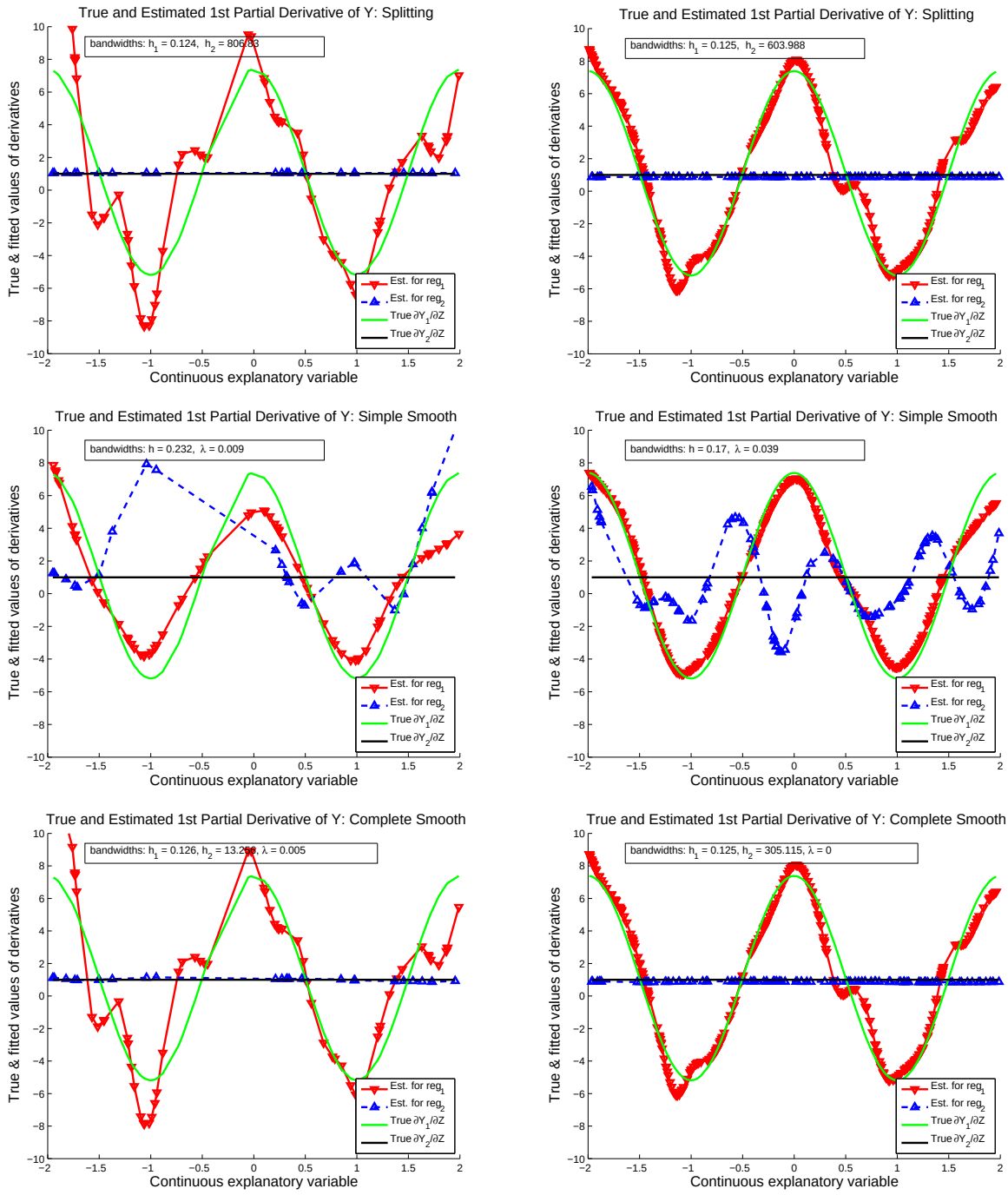

Figure 6: *Derivative Estimates for Example 1: Left panel,* $n = 100$ *and right panel,* $n = 400$*. From top to bottom: Approach 1 (non-smoothing for the discrete variable), Approach 2 (local linear simple- smoothing), Approach 3 (local linear complete-smoothing).*

on average, higher than those of the richer countries, and thus the poorer countries would eventually catch up with or converge to the levels of GDP per capita of the richer countries.

This is often referred to as the (unconditional) "beta-convergence" phenomenon. Earlier works on this issue employed the parametric regression models. Some studies found that the slope coefficient (the "beta") in such regressions is negative and significantly different from zero, which supports the "beta-convergence" hypothesis. However, other studies found that the "beta" is insignificantly different from zero (i.e., no convergence) or even positive (i.e., "beta-divergence") and significantly different from zero for different samples or for distinct groups of countries within a sample or when additional explanatory variables are accounted for.[5] We next use the nonparametric regression method, which may give some useful insights. In particular, we apply the LLLS with the three approaches discussed above to the following regression relationship:

$$y_i = m(z_i^c, z_i^d) + \varepsilon_i,\ i = 1, \ldots, n$$

where y_i is growth in GDP per capita of country i between 1965 and 1990, z_i^c is the natural log of GDP per capita of country i in 1965, while z_i^d is a discrete variable which will be defined later and ε_i is a stationary noise satisfying $\mathrm{E}(\varepsilon_i|z_i^c, z_i^d) = 0$ and $\mathrm{Var}(\varepsilon_i|z_i^c, z_i^d) < \infty$ a.s.

The estimation results are shown in Figure 7 which also includes some information on the resulting bandwidths chosen by different approaches. In panel (a) of Figure 7, we give the LLLS estimated curve without the discrete variable with h chosen by the CV method. From this figure, one may conjecture that there might be different groups within the sample which have different regression relationships (in terms of intercept or slope or both). Indeed, in various existing studies, researchers often distinguish various groups of countries, allowing them to have different regression relationships. An objective grouping criterion often used in practice, for example, is an indicator whether a country is OECD member or not (e.g., Racine et al., 2006; Simar and Zelenyuk, 2006; Maasoumi et al., 2007; Henderson and Zelenyuk, 2007), and so we also use this as our discrete variable, z_i^d, that has value 1 if country i was a member of OECD in the year 1965 and zero otherwise.[6]

In panel (b) of Figure 7, we give the LLLS estimated curves by using Approach 1 (i.e., separate estimation for each group with h chosen via CV for each group separately), and one can see that the estimated relationships for the two groups are very different not only in the intercept but also in the slope. Specifically, note that the relationship for the larger non-OECD group is virtually flat, with very slight inverted-U-shape curvature. On the other hand, note that the relationship for the smaller OECD group has a more pronounced

[5]For a recent review of this topic, see, for example, Maasoumi et al. (2007), Weil (2008) and references cited therein.

[6]In a detailed analysis, one may want to condition for many other potentially important explanatory variables, yet we will limit our illustration to the case of one continuous and one discrete explanatory variable for the sake of ease of graphical representation of the phenomenon.

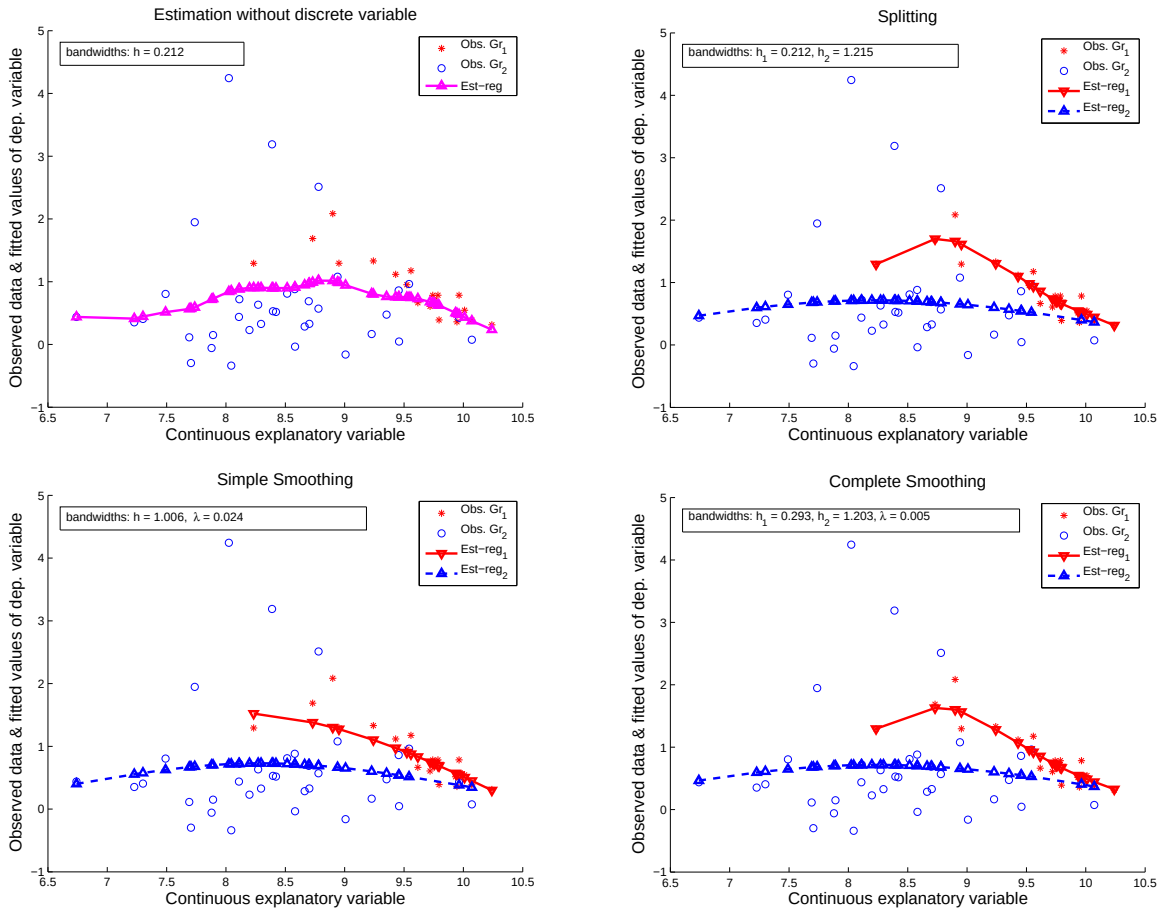

Figure 7: *Illustration with GDP data. From left to right and top to bottom, Panel (a): Approach 0 (without the discrete variable), Panel (b): Approach 1 (non-smoothing for the discrete variable), Panel (c): Approach 2 (local linear simple-smoothing), Panel (d): Approach 3 (local linear complete-smoothing).*

inverted-U-shape curvature (or rather "inverted hockey-stick" shape). Note that such curvature may suggest an important economic implication. It hints that the OECD countries with very low initial GDP per capita are expected to have higher growth rates in GDP per capita than those with very high initial GDP per capita, yet the highest rates are expected to be not at the lowest level of GDP per capita but somewhat larger.

In panel (c) of Figure 7, we present the LLLS estimated curves by using Approach 2 (i.e., local linear simple-smoothing with h and λ selected jointly via the CV method for the entire sample). One can see that the estimated relationships are also very different between the two groups. The relationship for the non-OECD group has slightly more pronounced inverted-U-shape curvature, although it still remains relatively flat. On the other hand, the relationship for the OECD group has much less curvature than that in panel (b), and it is

not an inverted-U-shape at all. Hence, Approach 2 suggests that for the OECD countries there is almost linear and negative relationship between the growth in GDP per capita and the initial level of GDP per capita. In other words, with the local linear simple-smoothing approach, we get some under-smoothing for the larger group and over-smoothing for the smaller group compared with panel (b) by separate estimation approach. This is similar to what we have observed from the simulated examples.

Some additional insight is provided by the local linear complete-smoothing method (Approach 3), where we allow for each group identified by the discrete variable to have its own bandwidth but also smooth the discrete variable and so use the full sample in one estimation. Panel (d) of Figure 7 visualizes the estimated curves by using Approach 3 and one can see that it gives almost identical results to those by using Approach 1, which is similar to what we have observed from the simulation studies. In this small data set, we can also find that the left-most observation in group 1 might be outlier, and so omitting it when using Approaches 1 and 3 may produce results very similar to Approach 2. However, it might be the case that there are other data points not available in our sample which are close to this left-most observation in group 1 and including them would make the inverted U-shape curvature even more pronounced. Since we do not know the true relationship, unlike in the simulated examples, it is difficult to judge which of these two arguments is likely to be right or wrong, which is beyond the scope of this paper. Since Approach 3 encompasses the other two approaches by taking the best features from each, and that our simulations suggested that Approach 3 was more robust than the other two, Approach 3 appears to be more reliable for a practitioner to trust in this context and perhaps in general, whenever it is computationally feasible.

Finally, it might be worth emphasizing again that in this section we had not intended to resolve the puzzles of economic growth across countries as such study would require larger data set and more variables. Our aim was just to give a concise and vivid illustration of the phenomenon we discussed above and, in particular, to compare the three approaches, not only for simulated data sets, but also for a real data, in a context that appears to be interesting for a wide audience.

6 Conclusion

In this article we have pointed out and illustrated that the reduction in variance or the efficiency gain due to smoothing of the discrete regressors with common bandwidth for the continuous variables across groups, as is frequently done in applied studies, can be well outweighed by the substantial bias introduced due to this simple-smoothing approach, both for small and for relatively large sample cases. For such cases, even fully separate estimation

for each group, if feasible, might be preferred. We have shown that the complete-smoothing technique by allowing different bandwidths for the continuous variables in each group, could overcome this difficulty and so is more robust than the existing smoothing methods. In general, whether it is better to smooth or not to smooth the discrete variable, or whether "the bias beats the variance", essentially depends on the degree of difference of the DGPs in different groups: curvatures of the regression relationship, variation in the error term for each group, variation in the continuous regressors, size or proportion of one group relative to another in the sample, and so on. The more robust complete-smoothing approach proposed in this paper is indeed a generalization of the seminal work by Racine and Li (2004), but at a cost of slightly more computational complexity

When using the simple-smoothing method, one automatically (or implicitly) imposes the assumption of similar degree of smoothness of the regression relationships for different groups of the sample identified by the discrete variables, which might be far from reality. As we illustrated in both simulated and empirical examples, such restriction can significantly deteriorate estimation results, increasing bias in the estimates of the true regression relationship. Such problem can also substantially or even radically distort estimates of derivatives and the related estimates of marginal effects and elasticities, which are used to draw policy implications.

It is also important to recognize that even from a theoretical point of view, the simple-smoothing of the discrete variable is preferable and offers a suitable solution to the problem, it is still an open question in practice for some real data sets, whether we have to smooth or not to smooth the discrete variables, particularly when the computational cost of the extended method is prohibitive. Further theoretical work is thus needed to develop and justify a method (specification test or a rule of thumb) that would help justifying a decision whether to smooth or not to smooth over some or all discrete variables. The issue of relevance of some categorical predictors in nonparametric regressions has been analyzed in Racine et al. (2006) by considering the hypothesis testing problem of $\lambda = 1$. However, to the best of our knowledge, nothing has been done for the other extreme of the scale ($\lambda = 0$), including the issue of common bandwidths for the continuous variables. Another possible extension of this paper would be to investigate whether the complete-smoothing method we proposed in this paper can also improve the performance of various tests that employ simple-smoothing method (Racine et al., 2006; Hsiao et al., 2007).

7 Acknowledgements

The authors would like to thank Peter Phillips and Jeff Racine for the inspirational comments that help improving the paper from earlier versions. Thanks also go to the colleagues

who commented on this paper and participants of various conferences and seminars where this work was presented. The first author acknowledges the financial support from ARC Discovery Early Career Researcher Award (DE120101130) and Monash Researcher Accelerator Plan. The second author and the third author acknowledge the financial support from ARC Discovery Grant (DP130101022), the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) and the School of Economics and CEPA of The University of Queensland. Only the authors and not the above mentioned institutions or people remain responsible for the views expressed.

A Assumptions

In this appendix, we give the regularity assumptions which are sufficient to derive the asymptotic theory of the proposed approach.

ASSUMPTION 1. Let $\{(Y_i, Z_i^c, Z_i^d)\}$ be independent and identically distributed (i.i.d.) as (Y, Z^c, Z^d), and the error term ε_i have the $(2+\delta)$ moment with $\delta > 0$.

ASSUMPTION 2. Let $K(\cdot)$ be a continuous and symmetric probability density function with a compact support.

ASSUMPTION 3. The conditional density function of Z^c for given Z^d, $f(z^c|z^d)$, is bounded away from infinity and zero for $z^c \in \mathcal{Z}^c$ and $z^d = z^d(1)$ or $z^d(2)$, where $\mathcal{Z}^c$ is the compact support of Z^c. Meanwhile, $m(\cdot, z^d)$, $\sigma^2(\cdot, z^d)$ and $f(\cdot|z^d)$ are twice continuously differentiable on $\mathcal{Z}^c$ for $z^d = z^d(1)$ or $z^d(2)$, where $\sigma^2(z^c, z^d) = \mathrm{Var}\big[\varepsilon|Z = (z^c, z^d)\big]$.

ASSUMPTION 4. Let the bandwidths $h(1)$, $h(2)$ and λ satisfy

$$h(1) \vee h(2) \to 0, \quad \big[nh(1)\big] \wedge \big[nh(2)\big] \to \infty \ \text{ and } \ \lambda = O(h^2(1) \wedge h^2(2)).$$

ASSUMPTION 4′. Let the bandwidths $h(1)$ and $h(2)$ satisfy

$$\big[n^{2\epsilon-1}h(1)\big] \wedge \big[n^{2\epsilon-1}h(2)\big] \to \infty, \quad \epsilon < (\delta+1)/(2+\delta),$$

where δ is defined in Assumption 1.

In Assumption 1, we impose the i.i.d. condition on the observations, which has been widely used in the literature on nonparametric estimation with both categorical and continuous data, see, for example, Li and Racine (2004) and Racine and Li (2004). We conjecture that our asymptotic theory can be generalized to some stationary and weakly dependent (such as β-mixing dependent) processes at the cost of more lengthy proofs. Assumption

2 imposes some mild conditions on the kernel function, and several commonly-used kernel functions such as the uniform kernel and the Epanechnikov kernel satisfy these conditions. Assumption 3 imposes some smoothness conditions on the conditional density function, conditional regression function and conditional variance function, which are necessary when the LLLS estimation approach is applied. Assumptions 4 and 4′ impose some restrictions on the bandwidths. In particular, Assumption 4′ is critical to apply the uniform consistency results of the nonparametric kernel estimators.

B Proofs of the Asymptotic Results

We next give the proofs of the asymptotic results stated in Section 3.

Proof of Theorem 3.1. Let

$$\begin{aligned}
\Delta_n(1) &= \sum_{i=1}^{n} \begin{bmatrix} 1 & (Z_i^c - z^c) \\ (Z_i^c - z^c) & (Z_i^c - z^c)^2 \end{bmatrix} \Lambda_\lambda(Z_i^d, z^d) K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\}, \\
\Delta_n(2) &= \sum_{i=1}^{n} \begin{bmatrix} 1 & (Z_i^c - z^c) \\ (Z_i^c - z^c) & (Z_i^c - z^c)^2 \end{bmatrix} \Lambda_\lambda(Z_i^d, z^d) K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\}, \\
\Omega_n(1) &= \sum_{i=1}^{n} \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \widetilde{Y}_i \Lambda_\lambda(Z_i^d, z^d) K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\}, \\
\Omega_n(2) &= \sum_{i=1}^{n} \begin{pmatrix} 1 \\ (Z_i^c - z^c) \end{pmatrix} \widetilde{Y}_i \Lambda_\lambda(Z_i^d, z^d) K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\},
\end{aligned}$$

where $\widetilde{Y}_i = Y_i - m(z^c, z^d) - m'(z^c, z^d)(Z_i^c - z^c)$, $m'(z^c, z^d)$ is the first-order partial derivative of $m(\cdot, \cdot)$ with respect to z^c, and let $e_2(1)$ be a 2-dimensional column vector with the first element being 1 and elsewhere 0.

By some elementary calculations, we can show that

$$\zeta(z^c, z^d) = \widetilde{m}(z^c, z^d) - m(z^c, z^d) = e_2^\tau(1) \Delta_n^+ \Omega_n, \tag{B.1}$$

where $\Delta_n = \Delta_n(1) + \Delta_n(2)$, $\Omega_n = \Omega_n(1) + \Omega_n(2)$ and Δ_n^+ is the Moore-Penrose inverse matrix of Δ_n.

We first consider the case of $z^d = z^d(1)$. Note that for this case,

$$\Delta_n(1) = \sum_{i=1}^{n} \begin{bmatrix} 1 & (Z_i^c - z^c) \\ (Z_i^c - z^c) & (Z_i^c - z^c)^2 \end{bmatrix} K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\}$$

and

$$\Delta_n(1) = \mathrm{E}[\Delta_n(1)] + \Delta_n(1) - \mathrm{E}[\Delta_n(1)].$$

By Assumptions 2–4 and standard argument, we can prove

$$\frac{1}{n}H_1^+\mathrm{E}[\Delta_n(1)]H_1^+ = p_1 f(z^c|z^d(1))\Delta(K) + o_P(1), \tag{B.2}$$

where $H_1 = \mathrm{diag}(1, h(1))$, $p_1 = P(Z^d = z^d(1))$, and $\Delta(K) = \mathrm{diag}(1, \mu_2)$. Furthermore, we can show that

$$\mathrm{Var}\Big[\frac{1}{n}H_1^+\Delta_n(1)H_1^+\Big] = O\big(\frac{1}{nh(1)}\big) = o(1)$$

as $nh_1 \to \infty$, which implies that

$$\frac{1}{n}H_1^+\Big\{\Delta_n(1) - \mathrm{E}[\Delta_n(1)]\Big\}H_1^+ = o_P(1). \tag{B.3}$$

Equations (B.2) and (B.3) lead to

$$\frac{1}{n}H_1^+\Delta_n(1)H_1^+ = p_1 f(z^c|z^d(1))\Delta(K) + o_P(1). \tag{B.4}$$

On the other hand, when $z^d = z^d(1)$, note that

$$\Delta_n(2) \;=\; \lambda\sum_{i=1}^n \begin{bmatrix} 1 & (Z_i^c - z^c) \\ (Z_i^c - z^c) & (Z_i^c - z^c)^2 \end{bmatrix} K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\}.$$

By the condition $\lambda = O(h^2(1) \wedge h^2(2))$ in Assumption 4, following the proof of (B.4), $\Delta_n(2)$ is dominated by $\Delta_n(1)$, which implies that

$$\frac{1}{n}H_1^+\Delta_n H_1^+ = \frac{1}{n}H_1^+\Delta_n(1)H_1^+ + o_P(1) = p_1 f(z^c|z^d(1))\Delta(K) + o_P(1). \tag{B.5}$$

We next look into Ω_n for the case of $z^d = z^d(1)$. Observe that

$$\begin{aligned}
\Omega_n(1) &= \sum_{i=1}^n \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \widetilde{Y}_i K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\}, \\
\Omega_n(2) &= \lambda\sum_{i=1}^n \begin{pmatrix} 1 \\ (Z_i^c - z^c) \end{pmatrix} \widetilde{Y}_i K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\}.
\end{aligned}$$

Furthermore, by the definition of the model, we can show that

$$\begin{aligned}
\Omega_n(1) &= \sum_{i=1}^n \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \rho_1(Z_i^c) K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\} \\
&\quad + \sum_{i=1}^n \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \varepsilon_i K_{h(1)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(1)\big\} \\
&=: \Omega_n(1,1) + \Omega_n(1,2),
\end{aligned}$$

where $\rho_1(Z_i^c) = m(Z_i^c, z^d(1)) - m(z^c, z^d(1)) - m'(z^c, z^d(1))(Z_i^c - z^c)$, and

$$\begin{aligned}\Omega_n(2) &= \lambda \sum_{i=1}^{n} \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \rho_2(Z_i^c) K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\} \\ &\quad + \lambda \sum_{i=1}^{n} \begin{pmatrix} 1 \\ Z_i^c - z^c \end{pmatrix} \varepsilon_i K_{h(2)}(Z_i^c - z^c) I\big\{Z_i^d = z^d(2)\big\} \\ &=: \Omega_n(2,1) + \Omega_n(2,2),\end{aligned}$$

where $\rho_2(Z_i^c) = m(Z_i^c, z^d(2)) - m(z^c, z^d(1)) - m'(z^c, z^d(1))(Z_i^c - z^c)$. As $\mathrm{E}(\varepsilon|Z) = 0$ a.s., $\Omega_n(1,1)$ and $\Omega_n(2,1)$ contribute to the asymptotic bias of the local linear estimator with complete smoothing. By standard argument for the local linear smoothing, we can show that

$$\frac{1}{n} H_1^+ \Omega_n(1,1) = \frac{1}{2} p_1 f(z^c|z^d(1)) \mu_2 m''(z^c, z^d(1)) h^2(1) e_2(1) + O_P(h^4(1)). \tag{B.6}$$

On the other hand, notice that

$$\rho_2(Z_i^c) = m(Z_i^c, z^d(2)) - m(z^c, z^d(1)) - m'(z^c, z^d(1))(Z_i^c - z^c) \approx m(z^c, z^d(2)) - m(z^c, z^d(1)).$$

We can thus prove that

$$\frac{1}{n} H_1^+ \Omega_n(2,1) = \lambda(1-p_1) f(z^c|z^d(1)) \big[m(z^c, z^d(2)) - m(z^c, z^d(1))\big] + o_P(\lambda). \tag{B.7}$$

As an additional factor λ is involved in $\Omega_n(2,2)$ and $\lambda = O(h^2(2))$, $\Omega_n(2,2)$ would be dominated by $\Omega_n(1,2)$. Hence, we next only need to prove the central limit theorem for $\Omega_n(1,2)$. Note that $\Omega_n(1,2)$ is a sum of independent and identically distributed (i.i.d.) random vectors with

$$\mathrm{E}\big[\Omega_n(1,2)\big] = 0$$

and

$$\frac{1}{nh(1)} \mathrm{Var}\big[H_1^+ \Omega_n(1,2)\big] = p_1 f(z^c|z^d(1)) \sigma^2(z^c, z^d(1)) \Omega(K),$$

where $\Omega(K) = \mathrm{diag}(\nu_0, \nu_2)$. By the classical central limit theorem and noting that

$$\widetilde{m}(z^c, z^d) - m(z^c, z^d) = e_2^\tau(1)(H_1 \Delta_n^+ H_1)(H_1^+ \Omega_n), \tag{B.8}$$

we can complete the proof of (3.2).

The proof of Theorem 3.1 for the case of $z^d = z^d(2)$ is similar and thus the details are omitted here.

Proof of Theorem 3.2. To simplify the notations, we let

$$\widetilde{m}_{(-i)}(Z_i^c, Z_i^d) = \widetilde{m}_{(-i)}(Z_i^c, Z_i^d \big| h(1), h(2), \lambda),$$

$\zeta_{(-i)}(Z_i^c, Z_i^d) = \widetilde{m}_{(-i)}(Z_i^c, Z_i^d) - m(Z_i^c, Z_i^d)$ and $f_Z(z^c, z^d) = f(z^c|z^d)P(Z^d = z^d)$.

Note that

$$\begin{aligned}
\overline{\mathrm{CV}}(h(1), h(2), \lambda) &= \sum_{i=1}^{n} \left[\varepsilon_i + m(Z_i^c, Z_i^d) - \widetilde{m}_{(-i)}(Z_i^c, Z_i^d)\right]^2 M(Z_i^c, Z_i^d) \\
&= \sum_{i=1}^{n} \varepsilon_i^2 M(Z_i^c, Z_i^d) - 2\sum_{i=1}^{n} \varepsilon_i \zeta_{(-i)}(Z_i^c, Z_i^d) M(Z_i^c, Z_i^d) \\
&\quad + \sum_{i=1}^{n} \zeta_{(-i)}^2(Z_i^c, Z_i^d) M(Z_i^c, Z_i^d) \\
&=: \overline{\mathrm{CV}}_1 + \overline{\mathrm{CV}}_2(h(1), h(2), \lambda) + \overline{\mathrm{CV}}_3(h(1), h(2), \lambda). \qquad \text{(B.9)}
\end{aligned}$$

It is easy to see that $\overline{\mathrm{CV}}_1 = \sum_{i=1}^{n} \varepsilon_i^2 M(Z_i^c, Z_i^d)$ does not rely on the bandwidths $h(1)$, $h(2)$ and λ, which implies that it would not play any role in choosing the optimal bandwidths.

We next derive the asymptotic order for $\overline{\mathrm{CV}}_2(h(1), h(2), \lambda)$. Define

$$\begin{aligned}
\Delta_{(-i)}(1) &= \sum_{j=1,\neq i}^{n} \begin{bmatrix} 1 & (Z_j^c - Z_i^c) \\ (Z_j^c - Z_i^c) & (Z_j^c - Z_i^c)^2 \end{bmatrix} \Lambda_\lambda(Z_j^d, Z_i^d) K_{h(1)}(Z_j^c - Z_i^c) I\{Z_j^d = z^d(1)\}, \\
\Delta_{(-i)}(2) &= \sum_{j=1,\neq i}^{n} \begin{bmatrix} 1 & (Z_j^c - Z_i^c) \\ (Z_j^c - Z_i^c) & (Z_j^c - Z_i^c)^2 \end{bmatrix} \Lambda_\lambda(Z_j^d, Z_i^d) K_{h(2)}(Z_j^c - Z_i^c) I\{Z_j^d = z^d(2)\}, \\
\Omega_{(-i)}(1) &= \sum_{j=1,\neq i}^{n} \begin{pmatrix} 1 \\ Z_j^c - Z_i^c \end{pmatrix} \widetilde{Y}_j \Lambda_\lambda(Z_j^d, Z_i^d) K_{h(1)}(Z_j^c - Z_i^c) I\{Z_j^d = z^d(1)\}, \\
\Omega_{(-i)}(2) &= \sum_{j=1,\neq i}^{n} \begin{pmatrix} 1 \\ (Z_j^c - Z_i^c) \end{pmatrix} \widetilde{Y}_i \Lambda_\lambda(Z_j^d, Z_i^d) K_{h(2)}(Z_j^c - Z_i^c) I\{Z_j^d = z^d(2)\}
\end{aligned}$$

and $\Delta_{(-i)} = \Delta_{(-i)}(1) + \Delta_{(-i)}(2)$, $\Omega_{(-i)} = \Omega_{(-i)}(1) + \Omega_{(-i)}(2)$. It is easy to show that

$$\zeta_{(-i)}(Z_i^c, Z_i^d) = e_2^\tau(1) \Delta_{(-i)}^+ \Omega_{(-i)}. \qquad \text{(B.10)}$$

By Lemma C.1 in Appendix C, we can prove that

$$\sup_{1 \le i \le n} \left| \frac{1}{n} H^+ \Delta_{(-i)} H^+ - f_Z(Z_i^c, Z_i^d) \Delta(K) \right| = o_P(1). \qquad \text{(B.11)}$$

Furthermore, let $\Omega_{(-i)} = \Omega_{(-i)}(1,1) + \Omega_{(-i)}(1,2) + \Omega_{(-i)}(2,1) + \Omega_{(-i)}(2,2)$, where $\Omega_{(-i)}(j_1, j_2)$ are defined as $\Omega_n(j_1, j_2)$ in the proof of Theorem 3.1 with the i-th observation (Y_i, Z_i) left out in the summation. By Assumption 4, (B.11) and Lemma C.1 again, we prove that

$$\sum_{i=1}^{n} \varepsilon_i e_2^\tau(1) \Delta_{(-i)}^+ \left[\Omega_{(-i)}(1,1) + \Omega_{(-i)}(2,1)\right] = O_P\left(\sqrt{n} h^2(1) + \sqrt{n} h^2(2)\right). \qquad \text{(B.12)}$$

By Lemma C.2 in Appendix C, we have

$$\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n}\varepsilon_i K\Big(\frac{Z_j^c - Z_i^c}{h(k)}\Big)\varepsilon_j = O_P\big(n\sqrt{h(k)}\big), \quad k=1,2, \tag{B.13}$$

which together with (B.11), leads to

$$\sum_{i=1}^{n}\varepsilon_i e_2^\tau(1)\Delta_{(-i)}^{+}\big[\Omega_{(-i)}(1,2)+\Omega_{(-i)}(2,2)\big] = O_P\big(\frac{1}{\sqrt{h(1)}}+\frac{1}{\sqrt{h(2)}}\big). \tag{B.14}$$

By (B.12) and (B.14), we have

$$\overline{\mathrm{CV}}_2(h(1),h(2),\lambda) = O_P\big(\sqrt{n}h^2(1)+\sqrt{n}h^2(2)+\frac{1}{\sqrt{h(1)}}+\frac{1}{\sqrt{h(2)}}\big). \tag{B.15}$$

We next consider $\overline{\mathrm{CV}}_3(h(1),h(2),\lambda)$ and prove that it can dominate $\overline{\mathrm{CV}}_2(h(1),h(2),\lambda)$. Let $\Omega_{(-i)}(\cdot,1)=\Omega_{(-i)}(1,1)+\Omega_{(-i)}(2,1)$ and $\Omega_{(-i)}(\cdot,2)=\Omega_{(-i)}(1,2)+\Omega_{(-i)}(2,2)$. As

$$\zeta_{(-i)}(Z_i^c,Z_i^d) = e_2^\tau(1)\Delta_{(-i)}^{+}\big[\Omega_{(-i)}(\cdot,1)+\Omega_{(-i)}(\cdot,2)\big],$$

we can show that

$$\begin{aligned}
&\sum_{i=1}^{n}\zeta_{(-i)}^2(Z_i^c,Z_i^d)M(Z_i^c,Z_i^d)\\
=\;&\sum_{i=1}^{n}\big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,1)\big]^2M(Z_i^c,Z_i^d)+\sum_{i=1}^{n}\big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,2)\big]^2M(Z_i^c,Z_i^d)\\
&+2\sum_{i=1}^{n}\big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,1)\big]\cdot\big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,2)\big]M(Z_i^c,Z_i^d)\\
=:\;&\overline{\mathrm{CV}}_{31}(h(1),h(2),\lambda)+\overline{\mathrm{CV}}_{32}(h(1),h(2),\lambda)+\overline{\mathrm{CV}}_{33}(h(1),h(2),\lambda).
\end{aligned}\tag{B.16}$$

By Taylor's expansion for $m(\cdot,\cdot)$ and using the uniform consistency results such as Lemma C.1, we can prove that

$$\sup_{1\le i\le n}\Big|e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,1)-b_m(Z_i^c,h(1),h(2),\lambda)\Big| = O_P\big(h^4(1)+h^4(2)\big), \tag{B.17}$$

where

$$b_m(Z_i^c,h(1),h(2),\lambda) = \begin{cases} b_*(Z_i^c,z^d(1))h^2(1)+\lambda(1-p_1)\big[m(Z_i^c,z^d(2))-m(Z_i^c,z^d(1))\big]/p_1, & Z_i^d=z^d(1),\\ b_*(Z_i^c,z^d(2))h^2(2)+\lambda p_1\big[m(Z_i^c,z^d(1))-m(Z_i^c,z^d(2))\big]/(1-p_1), & Z_i^d=z^d(2),\end{cases}$$

and $b_*(\cdot,\cdot)$ is defined in Section 3. Then, by (B.17) and the Law of Large Numbers, we can

show that

$$\begin{aligned}\overline{\mathrm{CV}}_{31}(h(1),h(2),\lambda) &= \sum_{i=1}^{n} b_m^2(Z_i^c,h(1),h(2),\lambda)M(Z_i^c,Z_i^d) \\ &= \sum_{i=1}^{n} b_m^2(Z_i^c,h(1),h(2),\lambda)M(Z_i^c,Z_i^d)I\big\{Z_i^d = z^d(1)\big\} \\ &\quad + \sum_{i=1}^{n} b_m^2(Z_i^c,h(1),h(2),\lambda)M(Z_i^c,Z_i^d)I\big\{Z_i^d = z^d(1)\big\} \\ &= n\big[\psi_1(h(1),\lambda)+\psi_2(h(2),\lambda)\big] + s.o., \end{aligned} \tag{B.18}$$

where $\psi_1(h(1),\lambda)$ and $\psi_2(h(2),\lambda)$ are defined in Section 3. By Lemma C.3 in Appendix C and some tedious calculations, we can prove that

$$\begin{aligned}\overline{\mathrm{CV}}_{32}(h(1),h(2),\lambda) &= \sum_{i=1}^{n} \big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,2)\big]^2 M(Z_i^c,Z_i^d)I\big\{Z_i^d = z^d(1)\big\} \\ &\quad \sum_{i=1}^{n} \big[e_2^\tau(1)\Delta_{(-i)}^{+}\Omega_{(-i)}(\cdot,2)\big]^2 M(Z_i^c,Z_i^d)I\big\{Z_i^d = z^d(2)\big\} \\ &= \chi(h(1))+\chi(h(2))+s.o., \end{aligned} \tag{B.19}$$

where $\chi(h(1))$ and $\chi(h(2))$ are defined in Section 3. Furthermore, by some standard but tedious calculations, we can also show that $\overline{\mathrm{CV}}_{33}(h(1),h(2),\lambda)$ is dominated by $\overline{\mathrm{CV}}_{31}(h(1),h(2),\lambda)+\overline{\mathrm{CV}}_{32}(h(1),h(2),\lambda)$. Hence, we have

$$\overline{\mathrm{CV}}_{3}(h(1),h(2),\lambda) = n\big[\psi_1(h(1),\lambda)+\psi_2(h(2),\lambda)\big] + \chi(h(1))+\chi(h(2))+s.o. \tag{B.20}$$

Then, by (B.9), (B.15), (B.20) and Assumption 4, we can complete the proof of Theorem 3.2.

C Some Auxiliary Lemmas

In this appendix, we give some auxiliary lemmas which have been used to prove the main theoretical results in Appendix B. The first lemma follows from a result in Mack and Silverman (1982).

LEMMA C.1. *Let $(X_1,Y_1),\cdots,(X_n,Y_n)$ be i.i.d. random vectors where $\{Y_i\}$ is a real-valued random sequence. Assume further that $\mathrm{E}|Y_i|^s<\infty$ and $\sup_x\int|y|^s f(x,y)dy<\infty$, where $f(\cdot,\cdot)$ denotes the joint density function of (X_i,Y_i). Let $K(\cdot)$ be a bounded positive function with a compact support satisfying a Lipschitz condition. If, in addition, $n^{2\epsilon-1}h\to\infty$ for some*

$\epsilon < 1 - s^{-1}$, *then*

$$\sup_{x\in\mathcal{S}} \left| \frac{1}{n}\sum_{i=1}^{n} \{K_h(X_i - x)Y_i - \mathrm{E}\left[K_h(X_i - x)Y_i\right]\} \right| = O_P(\sqrt{\frac{\log h^{-1}}{nh}}),$$

where $\mathcal{S}$ *is a compact support and* $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$.

LEMMA C.2. *Let* $(X_1, u_1), \cdots, (X_n, u_n)$ *be i.i.d. random vectors with*

$$\mathrm{E}(u_1|X_1) = 0 \quad \textit{and} \quad 0 < \mathrm{E}(u_i^2|X_i) < \infty \quad \textit{a.s.}$$

Assume further that $K(\cdot)$ *is a continuous and symmetric probability density function with a compact support,* h *is a bandwidth which tends to zero, and the density function of* X_i *is continuous and has a compact support. Then, we have*

$$\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n} u_i K\big(\frac{X_i - X_j}{h}\big)u_j = O_P(n\sqrt{h}). \tag{C.1}$$

PROOF OF LEMMA C.2. Letting $V_i = \sum_{j<i} K\big(\frac{X_i - X_j}{h}\big)u_j$, we have

$$\sum_{i=1}^{n}\sum_{j<i} u_i K\big(\frac{X_i - X_j}{h}\big)u_j = \sum_{i=1}^{n} u_i V_i.$$

It is easy to see that $\mathrm{E}(u_i V_i|\mathcal{F}_{i-1}) = 0$ a.s., where $\mathcal{F}_i = \sigma\big\{(X_k, u_k), k \le i-1, X_i\big\}$. Hence, $\{(u_i V_i, \mathcal{F}_i) : i \ge 1\}$ is a sequence of martingale differences. Using this fact, we can prove

$$\begin{aligned}
\mathrm{E}\Big[\big(\sum_{i=1}^{n}\sum_{j<i}^{n} u_i K\big(\frac{X_i - X_j}{h}\big)u_j\big)^2\Big] &= \mathrm{E}\Big[\big(\sum_{i=1}^{n} u_i V_i\big)^2\Big] = \sum_{i=1}^{n}\mathrm{E}[u_i^2 V_i^2] \\
&= \sum_{i=1}^{n}\mathrm{E}[u_i^2]\mathrm{E}\Big[\sum_{j<i} K^2\big(\frac{X_i - X_j}{h}\big)u_j^2\Big] \\
&= O(n^2 h).
\end{aligned} \tag{C.2}$$

Similarly, we can also prove

$$\mathrm{E}\Big[\big(\sum_{i=1}^{n}\sum_{j>i}^{n} u_i K\big(\frac{X_i - X_j}{h}\big)u_j\big)^2\Big] = O(n^2 h). \tag{C.3}$$

By (C.2), (C.3) and the Markov inequality, we can complete the proof of Lemma C.2.

LEMMA C.3. *Let the assumptions in Lemma C.2 are satisfied. Then, we have*

$$\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n}\sum_{k=1,\neq i}^{n} u_j K\big(\frac{X_i - X_j}{h}\big)K\big(\frac{X_i - X_k}{h}\big)u_k = V_* n^2 h + s.o. \tag{C.4}$$

where $V_* = \nu_0 \int \mathrm{E}[u_j^2|X_j = x]f^2(x)dx$ *and* $\nu_0 = \int K^2(x)dx$.

Proof of Lemma C.3. Note that

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n}\sum_{k=1,\neq i}^{n} u_j K\big(\frac{X_i - X_j}{h}\big)K\big(\frac{X_i - X_k}{h}\big)u_k \\
= &\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n} u_j^2 K^2\big(\frac{X_i - X_j}{h}\big) + \sum_{i=1}^{n}\sum_{j=1,\neq i}^{n}\sum_{k=1,\neq i,j}^{n} u_j K\big(\frac{X_i - X_j}{h}\big)K\big(\frac{X_i - X_k}{h}\big)u_k.
\end{aligned}$$

By Lemma C.1 and the Law of Large Numbers, we can prove

$$\begin{aligned}
\frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n} u_j^2 K^2\big(\frac{X_i - X_j}{h}\big) &= \frac{1}{n}\sum_{j=1}^{n} u_j^2\Big[\frac{1}{nh}\sum_{j=1,\neq i}^{n} K^2\big(\frac{X_i - X_j}{h}\big)\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n} u_j^2 f(X_j)\int K^2(x)dx \\
&= \nu_0 \int \mathrm{E}[u_j^2|X_j = x]f^2(x)dx. \qquad \text{(C.5)}
\end{aligned}$$

Similarly to the proof of Lemma C.2, we can show that

$$\sum_{i=1}^{n}\sum_{j=1,\neq i}^{n}\sum_{k=1,\neq i,j}^{n} u_j K\big(\frac{X_i - X_j}{h}\big)K\big(\frac{X_i - X_k}{h}\big)u_k = O_P(n^{3/2}h). \qquad \text{(C.6)}$$

The proof of (C.4) can be completed by using (C.5) and (C.6).

References

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

[2] Badunenko, O., D. Henderson and V. Zelenyuk (2008), Technological change and transition: relative contributions to worldwide growth during the 1990s, *Oxford Bulletin of Economics and Statistics*, 70, 461–492.

[3] Cleveland, W.S. (1979), Robust locally weighted regression and smoothing scatterplots, *Journal of American Statistical Association*, 74, 829–836.

[4] Cleveland, W.S. and S.J. Delvin (1988), Locally-weighted regression: an approach to regression analysis by local fitting, *Journal of American Statistical Association*, 83, 579–610.

[5] Eren, O. and D. Henderson (2008), The impact of homework on student achievement, *Econometrics Journal*, 11, 326–348.

[6] Fan, J. (1992), Design-adaptative nonparametric regression, *Journal of American Statistical Association*, 87, 998–1004.

[7] Fan, J. (1993), Local linear regression smoothers and their minimax efficiency, *Annals of Statistics*, 21, 196–216.

[8] Fan J. and I. Gijbels (1992), Variable bandwidth and local linear regression smoothers, *Annals of Statistics*, 20, 2008–2036.

[9] Fan J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall.

[10] Frölich, M. (2006), Non-parametric regression for binary dependent variables, *Econometrics Journal*, 9, 511–540.

[11] Gozalo, P. and O. Linton (2000), Local nonlinear least squares: Using parametric information in nonparametric regression, *Journal of Econometrics*, 99, 63–106.

[12] Hall, P., J.S. Racine and Q. Li (2004), Cross-Validation and the Estimation of Conditional Probability Densities, *Journal of the American Statistical Association*, Vol 99, 486, 1015–1026.

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

[14] Hartarska, V., C.F. Parmeter and D. Nadolynak (2010), Economies of scope of lending and mobilizing deposits in Microfinance institutions: A semiparametric Analysis, *American Journal of Agricultural Economics*, 93(2), 389–398.

[15] Henderson, D. (2010), A test for multimodality of regression derivatives with an application to nonparametric growth regressions, *Journal of Applied Econometrics*, 25, 458–480.

[16] Henderson, D.J. and R.R. Russell (2005), Human capital and convergence: a production-frontier approach, *International Economic Review*, Vol. 46, 1167–1205.

[17] Henderson, D.J. and V. Zelenyuk (2007), Testing for (efficiency) catching-up, *Southern Economic Journal*, Vol. 73, 1003–1019.

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

[19] Kumar, S. and R.R. Russell (2002), Technological change, technological catch-up, and capital deepening: Relative contributions to growth and convergence, *American Economic Review*, 92, 527–48.

[20] Li, Q. and J. Racine (2003), Nonparametric Estimation of Distributions With Categorical and Continuous Data, *Journal of Multivariate Analysis*, 86, 266-292.

[21] Li, Q. and J. Racine (2004), Cross-validated local linear nonparametric regression, *Statistica Sinica*, 14, 485–512.

[22] Li, Q. and J. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[23] Li, Q. and J. Racine (2008), Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data, *Journal of Business & Economic Statistics*, Vol 26 (4), 423–434.

[24] Li, Q., J. Racine and J.M. Wooldridge (2009), Efficient Estimation of Average Treatment Effect with Mixed Categorical and Continuous Data, *Journal of Business & Economic Statistics*, Vol 26 (4), 423–434.

[25] Maasoumi, E., J. Racine and T. Stengos (2007), Growth and convergence: A profile of distribution dynamics and mobility, *Journal of Econometrics*, 136, 483–508.

[26] Mack, Y.P. and B.W. Silverman (1982), Weak and strong uniform consistency for kernel regression estimates, *Zeitschrift fur Wahrscheinlichskeittheorie und verwandte Gebiete*, 61, 405–415.

[27] Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge University Press.

[28] Ouyang, D., Q. Li and J. Racine (2006), Cross-validation and the estimation of probability distributions with categorical data, *Nonparametric Statistics*, 18, 69–100.

[29] Park, B., L. Simar and V. Zelenyuk (2008), Local likelihood estimation of truncated regression and its partial derivatives: Theory and application, *Journal of Econometrics*, 146 (1), 185–198.

[30] Park, B., L. Simar and V. Zelenyuk (2010), Local Maximum Likelihood Methods with Categorical Variables. Discussion paper 1052, Institut de Statistique, UCL.

[31] Parmeter, C.F., D. Henderson and S.C. Kumbhakar (2007), Nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics*, 22, 695–699.

[32] Racine, J., J. Hart and Q. Li (2006), Testing the significance of categorical predictor variables in nonparametric regression models, *Econometric Review*, 25(4), 523–544.

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

[34] Ruppert, D. and M.P. Wand (1994), Multivariate weighted least squares regression, *Annals of Statistics*, 22, 1346–1370.

[35] Simar, L. and V. Zelenyuk (2006), On testing equality of two distribution functions of efficiency score estimated via DEA, *Econometric Reviews*, 25, 497–522.

[36] Stengos, T. and E. Zacharias (2006), Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market, *Journal of Applied Econometrics*, 21, 371–386.

[37] Stone, C. J. (1977), Consistent nonparametric regression, *Annals of Statistics*, 5, 595–645.

[38] Titterington, D. M. (1980), A comparative study of kernel-based density estimates for categorical data, *Technometrics*, 22, 259–268.

[39] Walls, W. (2009), Screen wars, star wars, and sequels, *Empirical Economics*, 37, 447–461.

[40] Wang, M.C. and J. Van Ryzin (1981), A class of smooth estimators for discrete distributions, *Biometrika*, 68, 301–309.

[41] Weil, D.N. (2008), *Economic Growth*, 2nd ed., Addison Wesley.

Table 1: Monte-Carlo Results for Example 1, over 500 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl
$\overline{AMSE}$ all	0.4556	0.6958	0.4707	0.2335	0.3889	0.2477	0.1230	0.2353	0.1305	0.0675	0.1353	0.0717
std_{MC} all	0.0123	0.0145	0.0119	0.0057	0.0069	0.0062	0.0031	0.0038	0.0034	0.0017	0.0022	0.0018
$\overline{AMSE}$ gr1	0.3053	0.4461	0.2999	0.1539	0.1907	0.1603	0.0837	0.1088	0.0853	0.0489	0.0628	0.0495
std_{MC} gr1	0.0103	0.0167	0.0081	0.0029	0.0041	0.0032	0.0015	0.0023	0.0016	0.0009	0.0011	0.0010
$\overline{AMSE}$ gr2	0.9631	1.5305	1.0371	0.4854	1.0114	0.5222	0.2440	0.6224	0.2713	0.1245	0.3540	0.1398
std_{MC} gr2	0.0410	0.0538	0.0434	0.0216	0.0298	0.0234	0.0121	0.0158	0.0137	0.0064	0.0083	0.0066
median $\hat{h}_1$	0.2062	0.2529	0.1884	0.1682	0.2115	0.1599	0.1424	0.1872	0.1378	0.1197	0.1615	0.1176
median $\hat{h}_2$	604.9	0.2529	2.532	3.174	0.2115	2.987	396.9	0.1872	4.05	177.5	0.1615	4.801
median $\hat{\lambda}$	0.0000	0.0770	0.0120	0.0000	0.0793	0.0059	0.0000	0.0545	0.0030	0.0000	0.0335	0.0022
median CV_{opt}	2.0404	2.1390	1.9020	1.9019	2.0199	1.8534	1.8206	1.9284	1.8094	1.7975	1.8618	1.7918

Table 2: Monte-Carlo Results for Example 2, over 500 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl
$\overline{AMSE}$ all	0.4556	0.6958	0.4707	0.2335	0.3889	0.2477	0.1230	0.2353	0.1305	0.0675	0.1353	0.0717
std_{MC} all	0.0123	0.0145	0.0119	0.0057	0.0069	0.0062	0.0031	0.0038	0.0034	0.0017	0.0022	0.0018
$\overline{AMSE}$ gr1	0.3053	0.4461	0.2999	0.1539	0.1907	0.1603	0.0837	0.1088	0.0853	0.0489	0.0628	0.0495
std_{MC} gr1	0.0103	0.0167	0.0081	0.0029	0.0041	0.0032	0.0015	0.0023	0.0016	0.0009	0.0011	0.0010
$\overline{AMSE}$ gr2	0.9631	1.5305	1.0371	0.4854	1.0114	0.5222	0.2440	0.6224	0.2713	0.1245	0.3540	0.1398
std_{MC} gr2	0.0410	0.0538	0.0434	0.0216	0.0298	0.0234	0.0121	0.0158	0.0137	0.0064	0.0083	0.0066
median $\hat{h}_1$	0.2062	0.2529	0.1884	0.1682	0.2115	0.1599	0.1424	0.1872	0.1378	0.1197	0.1615	0.1176
median $\hat{h}_2$	604.9	0.2529	2.532	3.174	0.2115	2.987	396.9	0.1872	4.05	177.5	0.1615	4.801
median $\hat{\lambda}$	0.0000	0.0770	0.0120	0.0000	0.0793	0.0059	0.0000	0.0545	0.0030	0.0000	0.0335	0.0022
median CV_{opt}	2.0404	2.1390	1.9020	1.9019	2.0199	1.8534	1.8206	1.9284	1.8094	1.7975	1.8618	1.7918

Table 3: Monte-Carlo Results for Example 3, over 500 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl
$\overline{AMSE}$ all	0.3145	0.2256	0.2470	0.1664	0.1220	0.1352	0.0848	0.0737	0.0804	0.0447	0.0428	0.0470
std_{MC} all	0.0102	0.0099	0.0109	0.0054	0.0048	0.0065	0.0031	0.0029	0.0053	0.0016	0.0012	0.0042
$\overline{AMSE}$ gr1	0.1140	0.1203	0.1240	0.0642	0.0734	0.0690	0.0326	0.0419	0.0358	0.0184	0.0242	0.0211
std_{MC} gr1	0.0040	0.0036	0.0044	0.0020	0.0021	0.0022	0.0011	0.0013	0.0012	0.0006	0.0007	0.0008
$\overline{AMSE}$ gr2	0.9631	0.5577	0.6354	0.4854	0.2774	0.3490	0.2440	0.1708	0.2152	0.1245	0.0994	0.1245
std_{MC} gr2	0.0410	0.0378	0.0420	0.0216	0.0198	0.0273	0.0121	0.0109	0.0205	0.0064	0.0044	0.0158
median $\hat{h}_1$	0.9742	1.1728	0.8906	0.7850	0.9226	0.7364	0.5964	0.7185	0.5840	0.5118	0.6185	0.5067
median $\hat{h}_2$	604.9	1.173	8.573	3.174	0.9226	5.892	396.9	0.7185	7.706	177.5	0.6185	6.115
median $\hat{\lambda}$	0.0000	0.3892	0.2901	0.0000	0.3324	0.1674	0.0000	0.3201	0.1262	0.0000	0.2342	0.0679
median CV_{opt}	1.8400	1.7030	1.6542	1.8171	1.7517	1.7207	1.7815	1.7515	1.7294	1.7725	1.7667	1.7568

Table 4: Monte-Carlo Results for Example 4, over 500 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl
$\overline{AMSE}$ all	0.3255	0.2805	0.3098	0.1708	0.1645	0.1685	0.0911	0.1004	0.0987	0.0491	0.0558	0.0499
std_{MC} all	0.0104	0.0094	0.0114	0.0053	0.0048	0.0056	0.0033	0.0032	0.0044	0.0016	0.0015	0.0015
$\overline{AMSE}$ gr1	0.1220	0.1388	0.1372	0.0675	0.0812	0.0720	0.0337	0.0457	0.0379	0.0192	0.0252	0.0210
std_{MC} gr1	0.0042	0.0041	0.0049	0.0022	0.0023	0.0023	0.0011	0.0014	0.0012	0.0006	0.0007	0.0007
$\overline{AMSE}$ gr2	0.9823	0.7313	0.8726	0.4926	0.4218	0.4677	0.2658	0.2673	0.2840	0.1399	0.1481	0.1370
std_{MC} gr2	0.0415	0.0362	0.0473	0.0209	0.0192	0.0226	0.0128	0.0118	0.0171	0.0062	0.0057	0.0056
median $\hat{h}_1$	0.8589	1.0182	0.7678	0.6942	0.8367	0.6530	0.5451	0.7069	0.4932	0.4701	0.6395	0.4325
median $\hat{h}_2$	319.8	1.018	3.822	2.588	0.8367	2.043	2.112	0.7069	1.428	1.999	0.6395	1.385
median $\hat{\lambda}$	0.0000	0.2256	0.1063	0.0000	0.1373	0.0532	0.0000	0.1059	0.0524	0.0000	0.0678	0.0313
median CV_{opt}	1.8432	1.7234	1.6830	1.8248	1.7877	1.7590	1.7824	1.7818	1.7652	1.7799	1.7820	1.7695

Table 5: Monte-Carlo Results for Example 5, over 500 MC replications.

	$n = 50$			$n = 100$			$n = 200$			$n = 400$		
col#	1	2	3	4	5	6	7	8	9	10	11	12
	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl	Split	Simpl	Compl
$\overline{AMSE}$ all	0.4547	0.6956	0.4740	0.2427	0.3871	0.2583	0.1342	0.2345	0.1396	0.0765	0.1352	0.0796
std_{MC} all	0.0114	0.0147	0.0122	0.0055	0.0070	0.0064	0.0033	0.0038	0.0035	0.0016	0.0022	0.0017
$\overline{AMSE}$ gr1	0.2951	0.4338	0.2978	0.1540	0.1923	0.1596	0.0837	0.1092	0.0857	0.0489	0.0630	0.0499
std_{MC} gr1	0.0089	0.0160	0.0086	0.0029	0.0041	0.0031	0.0015	0.0023	0.0016	0.0009	0.0011	0.0009
$\overline{AMSE}$ gr2	0.9938	1.5510	1.0549	0.5212	0.9992	0.5643	0.2889	0.6178	0.3059	0.1610	0.3528	0.1699
std_{MC} gr2	0.0410	0.0531	0.0442	0.0208	0.0296	0.0240	0.0127	0.0156	0.0140	0.0062	0.0083	0.0063
median $\hat{h}_1$	0.2068	0.2493	0.1837	0.1680	0.2112	0.1604	0.1423	0.1869	0.1375	0.1198	0.1614	0.1169
median $\hat{h}_2$	14.01	0.2493	2.201	2.172	0.2112	2.09	1.328	0.1869	1.25	1.135	0.1614	1.04
median $\hat{\lambda}$	0.0000	0.0809	0.0131	0.0000	0.0801	0.0081	0.0000	0.0549	0.0072	0.0000	0.0338	0.0056
median CV_{opt}	2.0511	2.1565	1.8884	1.9094	2.0193	1.8636	1.8332	1.9283	1.8183	1.8046	1.8618	1.8009