Electronic Journal of Statistics
Vol. 18 (2024) 803–847
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2216

High–dimensional local linear regression under sparsity and convex losses

Kin Yap Cheung and Stephen M.S. Lee

Department of Statistics & Actuarial Science
The University of Hong Kong
e-mail: u3507181@connect.hku.hk; smslee@hku.hk

Abstract: Existing works on variable selection and estimation for high-dimensional nonparametric regression focus primarily on modelling a conditional mean function, on which a restrictive additive structure is commonly imposed. We consider a more general framework which covers different types of regression derived from a broad class of convex loss functions, without assuming additivity of the nonparametric regression function to be estimated. A novel penalised local linear regression procedure is proposed for simultaneous variable selection and estimation under this framework. It performs Bridge-penalised local linear regression and regularised bandwidth estimation in a alternating optimisation scheme. The covariate dimension may exceed any polynomial order, while the number of active variables is allowed to grow slowly with sample size. The procedure is shown to be consistent in variable selection and yield a regression function estimator endowed with an oracle property. Simulation and real data examples are presented to illustrate the performance of the proposed method in mean regression, quantile regression and logistic regression problems.

MSC2020 subject classifications: Primary 62G08.
Keywords and phrases: Convex loss, high–dimensional, local linear regression, nonparametric regression, variable selection.

Received September 2023.

1. Introduction

Nonparametric regression is notoriously inefficient under high dimensions. Although many methods have been proposed for variable selection or structure simplification with the aim of improving convergence rates of nonparametric regression estimators, they are mostly conceived under fixed or slowly growing dimensions. Relatively little effort has been made in developing variable selection and regression methods under a general high-dimensional nonparametric framework. To fill this gap, we propose a new procedure for high-dimensional variable selection and regression under sparsity conditions and a general convex loss function, based on Bridge-penalised local linear estimation and bandwidth regularisation.

Methods for simultaneous variable selection and regression have been extensively studied under high-dimensional linear models, among which the LASSO is probably the most prominent. The need for variable selection is even more compelling in high-dimensional nonparametric regression problems, where the minimax convergence rate decreases exponentially with dimension. Two approaches, based respectively on projection and local polynomial fitting, are commonly employed for nonparametric regression. The projection approach, which embeds the regression function in a function space such as the reproducing kernel Hilbert space (RKHS), has been applied for variable selection under an additive model setting. Examples include different variants of the COSSO (Lin & Zhang, 2006; Zhang & Lin, 2006; Storlie et al., 2011; Lin et al., 2013), which are shown to work under fixed dimensions.

In the general case where additivity is not assumed, the existing variable selection methods are mainly confined to settings of fixed or slowly growing dimensions. For methods of the projection type, the RKHS remains a common choice for modelling the function space and has been considered by, for example, Ye & Xie (2012), Rosasco et al. (2013), Allen (2013), Yang et al. (2016) and Chen et al. (2017). Methods of the local polynomial type, designed mainly for selecting variables locally, include Bertin & Lecué (2008), Lafferty & Wasserman (2008), Miller & Hall (2010), Giordano & Parrella (2016) and White et al. (2017). Theoretical justification, if any, of the above methods is given under dimensions of at most a logarithmic order. An exception is Giordano et al. (2020), who propose a GRID method to identify relevant linear and nonlinear variables in multiple steps, under dimensions of a polynomial order. The method requires the regression function be dependent on interactions of a fixed order and the strong assumption that the covariates are uniformly distributed on the unit cube. Relatively few works have been devoted to explicit variable selection and model fitting under high dimensional and non-additive model settings. Aside from COSSO and Allen's (2013) KNIFE procedure, studies of the aforementioned methods are confined to a mean regression setting.

In the present paper, we propose a novel penalised local linear regression procedure in a sparse model setting. It allows applications to problem settings beyond mean regression and allows a sparsity level more general than previous methods. Our procedure performs variable selection and estimation simultaneously using local linear regression, and is able to handle a high dimension which may grow faster than any polynomial order of the sample size n. Variable selection is facilitated by a alternating optimisation scheme which regularises bandwidths coordinate-wise and regression coefficients in two separate steps. The output bandwidths help differentiate between active and inactive variables, as well as achieve an optimal rate of convergence in an oracle sense. Empirical results suggest that our method has better prediction performance than existing methods.

We state the problem in Section 2 and present our proposed procedure in Section 3. A coordinate descent algorithm is developed in Section 4 for the alternating optimisation scheme. We show that our method is selection consistent under very general conditions in Section 5. The asymptotic distribution of the regression function estimator is also derived there. Section 6 extends our method to quantile regression. Section 8 presents empirical studies to compare our procedure with existing methods. In the concluding Section 9, we summarise our results and discuss potential generalisations of our work.

2. Problem setting

We first introduce notations. For any $\mathcal{B} \subset \{1, \dots, D\}$ and $\boldsymbol{x} = (x_1, \dots, x_D)^\top \in \mathbb{R}^D$, denote by $|\mathcal{B}|$ the cardinality of $\mathcal{B}$ and by $\mathcal{B}^c$ the complement $\{1, \dots, D\} \backslash \mathcal{B}$, define $\|\boldsymbol{x}\|_q = \big(\sum_{d=1}^D |x_d|^q\big)^{1/q}$ for any $q > 0$, $\|\boldsymbol{x}\|_\infty = \max_{1 \le d \le D} |x_d|$ and $\boldsymbol{x}_\mathcal{B} \in \mathbb{R}^{|\mathcal{B}|}$ to be the column subvector of $\boldsymbol{x}$ formed by the components $\{x_d : d \in \mathcal{B}\}$. For any real-valued function g on an open neighbourhood of $\boldsymbol{x}$, define $\nabla_0 g(\boldsymbol{x}) = g(\boldsymbol{x})$, $\nabla_d g(\boldsymbol{x}) = \partial g(\boldsymbol{x})/\partial x_d$ and $\nabla_{d_1,d_2} g(\boldsymbol{x}) = \partial^2 g(\boldsymbol{x})/\partial x_{d_1} \partial x_{d_2}$, provided that the derivatives exist. For any $x, y \in \mathbb{R}$, define $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. The indicator function is denoted by $\mathbf{1}\{\cdot\}$.

Let $(\boldsymbol{X}, Y) = (X^{(1)}, \dots, X^{(D)}, Y)$ be a $(D+1)$-variate random vector drawn from an unknown distribution, with $Y \in \mathscr{Y}$ and $\boldsymbol{X}$ following a smooth density function f on $\mathbb{R}^D$. For $\boldsymbol{x} \in \mathbb{R}^D$, let $m(\boldsymbol{x})$ uniquely minimise $\mathbb{E}[L(Y, a)|\boldsymbol{X} = \boldsymbol{x}]$ with respect to $a \in \mathbb{R}$, for some loss function $L(Y, a)$ convex in a. We assume further that $\mathbb{E}[L(Y, a)|\boldsymbol{X} = \boldsymbol{x}]$ is strictly convex in a neighbourhood of $a = m(\boldsymbol{x})$. Our theory, established in Section 5, requires that partial derivatives of L be Lipschitz continuous: see (A6). The above conditions are satisfied by a rich class of nonparametric generalised linear models, under which $L(Y, a)$ is the negative loglikelihood derived from an exponential family of conditional log-densities

$$\xi(y|\boldsymbol{x}) = \xi_0(y) + \varphi^{-1}\big\{ym(\boldsymbol{x}) - b(m(\boldsymbol{x}))\big\}, \tag{2.1}$$

for some scale parameter $\varphi > 0$ and some function ξ_0 independent of $m(\boldsymbol{x})$. Common examples include normal, logistic and Poisson regression models. Local polynomial fitting for generalised linear models has been studied by Fan et al. (1995). The important problem of nonparametric quantile regression, for which (A6) fails to hold, is discussed separately in Section 6.

As to sparsity, we assume that there exists an index set $\mathcal{A} \subset \{1, \dots, D\}$ such that $m(\boldsymbol{x}) = m_\mathcal{A}(\boldsymbol{x}_\mathcal{A})$, $\boldsymbol{x} \in \mathbb{R}^D$, that is $m(\boldsymbol{x})$ depends only on a set of active variables $\boldsymbol{x}_\mathcal{A}$. The sparsity level $|\mathcal{A}|$ is allowed to grow slowly with n, while D may increase at a rate faster than any polynomial order.

Let $\mathscr{D}_n = \big((\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)\big)$ be n independent replicates of $(\boldsymbol{X}, Y)$, based on which the regression function $m(\boldsymbol{x})$ is to be estimated globally over $\boldsymbol{x} \in \mathbb{R}^D$. It is well known that the optimal error rate of a local linear regression estimator of $m(\boldsymbol{x})$ depends on the covariate dimension D. It can be substantially reduced to an oracle optimal rate if regression were done only on variables in $\mathcal{A}$ under sparsity, which underlines the importance of variable selection for high-dimensional regression problems. Our objective is to develop a local linear procedure capable of identifying $\mathcal{A}$ and estimating $m(\boldsymbol{x})$ at the oracle optimal rate. The results enable us to derive the asymptotic distribution of the regression function estimator, which suggests a straightforward procedure for inference about $m(\boldsymbol{x})$ at a fixed $\boldsymbol{x} \in \mathbb{R}^D$.

3. Methodology

In classical local polynomial regression, kernel bandwidths are required to converge to zero for consistent estimation of the regression function. Typically, the optimal convergence rates of the bandwidths decrease as the covariate dimension D increases, resulting in a slower convergence rate of the regression function estimator. Restricting attention to local constant regression, the MEKRO method proposed by White et al. (2017) minimises the mean squared residuals of a Nadaraya–Watson estimate, with bandwidths regularised by a constraint on the sum of their reciprocals. The estimated bandwidths provide a selection scheme whereby variables with large bandwidths are excluded and those with small bandwidths are kept in the model for consistent estimation of the regression function. We develop a local linear regression procedure, based on a alternating optimisation scheme, for variable selection and prediction at oracle optimal rates. Unlike theirs, our procedure applies to general convex loss functions under weaker sparsity conditions.

Let K be a non-negative, symmetric, bounded, kernel function on $\mathbb{R}$ such that $K(0) = \lim_{x\to 0} K(x) > 0$ and $\lim_{u\to\infty} uK(u) = 0$. Assume that, for $r \geq 1$ and $s \geq 0$, $\mu_{r,s} \equiv \int_{-\infty}^{\infty} u^s K(u)^r du$ exists and that $\mu_{1,0} = 1$ in particular. Note that symmetry of K implies $\mu_{r,s} = 0$ for all positive odd integer s. Define, for $\boldsymbol{h} = (h_1, \ldots, h_D)^\top \in (0, \infty]^D$ and $\boldsymbol{x} = (x_1, \ldots, x_D)^\top \in \mathbb{R}^D$, $K_{\boldsymbol{h}}(\boldsymbol{x}) = \prod_{d=1}^D h_d^{-1} K(x_d/h_d)$. The above are standard assumptions made on the kernel function used for local polynomial fitting.

Define, for $h > 0$,

$$C_n(h) = (nD)^{(\alpha\vee 2)-1} e^{-c/h} \quad \text{and} \quad \Lambda_n(h) = \lambda_n \left\{1 + \log\left(1 + h^4\right)\right\}^{-1},$$

where $c > 0$ and $\alpha \geq 1$ are fixed constants, and the positive sequence λ_n satisfies

$$\lambda_n^{-1} \omega^2 (n\omega^2)^{-4/(4+|\mathcal{A}|)} + \lambda_n |\mathcal{A}| = o(1),$$

for some constant $\omega \in [1, |\mathcal{A}|^2]$ to be specified in (A8). The sequence λ_n always exists provided that $|\mathcal{A}|$ satisfies the condition (B1) stated in Section 5. Under the slightly stronger condition that $|\mathcal{A}| \leq a_0 \log n / \log\log n$ for some $a_0 \in (0, 2/5)$, we may set $\lambda_n = 1/\log n$. Practical choices of the above tuning parameters in real applications will be discussed in Section 7. For any $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{h} \in (0, \infty]^D$ and $\mathcal{S} \subsetneq \{1, \ldots, n\}$, define

$$\begin{aligned}&\left(\hat{\beta}_0^{-\mathcal{S}}(\boldsymbol{x};\boldsymbol{h}), \ldots, \hat{\beta}_D^{-\mathcal{S}}(\boldsymbol{x};\boldsymbol{h})\right)^\top \\ &= \operatorname*{argmin}_{(\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^D} \left\{ \frac{\sum_{i\notin\mathcal{S}} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) L\left(Y_i, \beta_0 + \boldsymbol{\beta}^\top(\boldsymbol{X}_i - \boldsymbol{x})\right)}{\sum_{i\notin\mathcal{S}} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x})} + \sum_{d=1}^D C_n(h_d)|\beta_d|^\alpha \right\},\end{aligned} \tag{3.1}$$

which estimates the local linear coefficients $(\beta_0, \boldsymbol{\beta}^\top) = (\beta_0, \beta_1, \ldots, \beta_D)$ from a delete-$\mathcal{S}$ subsample based on a bandwidth-adaptive Bridge penalty. Let $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_K = \{1, \ldots, n\}$ be a partition of the index set with $K \geq 2$ and $|\mathcal{S}_k|/n + n/|\mathcal{S}_k| = O(1)$ for all $k = 1, \ldots, K$. Without loss of generality we assume $n_0 = |\mathcal{S}_1| =$

$\cdots = |\mathcal{S}_K|$. Define

$$\hat{\boldsymbol{h}} = (\hat{h}_1, \ldots, \hat{h}_D)^\top = \operatorname*{argmin}_{\boldsymbol{h} \in (0,\infty]^D} \Big\{ n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} L\big(Y_i, \hat{\beta}_0^{-\mathcal{S}_k}(\boldsymbol{X}_i; \boldsymbol{h})\big) + \sum_{d=1}^{D} \Lambda_n(h_d) \Big\} \tag{3.2}$$

to be a vector of regularised bandwidths obtained by minimising a delete-n_0 cross-validated loss, penalised by Λ_n and averaged over the given partition. Our proposed procedure essentially calculates $\hat{\boldsymbol{h}}$ by recursively evaluating (3.1) and (3.2) in a alternating optimisation algorithm. In (3.1), we associate to each local coefficient β_d a penalty $C_n(h_d)$ which decays exponentially fast with $1/h_d$, thereby suppressing those β_d's corresponding to large bandwidths while leaving free those corresponding to small bandwidths. Since consistent estimation of $m(\cdot)$ is possible only when bandwidths of nonlinear active variables are small, the cross-validated loss in (3.2) suppresses the growth of $\{h_d, d \in \mathcal{A}\}$, while the penalty $\Lambda_n(h_d)$ pushes $\{h_d, d \in \mathcal{A}^c\}$ towards infinity. We define $\Lambda_n(h_d)$ strategically to amplify the bandwidths $\hat{\boldsymbol{h}}_{\mathcal{A}^c}$ of inactive variables, while retaining optimality of the bandwidths $\hat{\boldsymbol{h}}_{\mathcal{A}}$ estimated for active variables. It is preferable to conventional penalties like L_p, for the latter lead to suboptimal bandwidths in general. For empirical determination of active variables, we may define the selected set $\hat{\mathcal{A}}$ to be $\{d : \hat{h}_d \le C_0\}$, for some threshold $C_0 > 0$ independent of n. With $\boldsymbol{h}$ duly estimated by $\hat{\boldsymbol{h}}$, we may estimate $m(\boldsymbol{x})$ by $\hat{m}(\boldsymbol{x}) = \hat{\beta}_0^{-\emptyset}(\boldsymbol{x}; \hat{\boldsymbol{h}})$. For a computationally more efficient alternative to $\hat{\beta}_0^{-\emptyset}(\boldsymbol{x}; \hat{\boldsymbol{h}})$, we may replace $\hat{h}_d$ by $\hat{h}_d/\mathbf{1}\{\hat{h}_d \le C_0\}$ in the calculation of $\hat{\boldsymbol{h}}$ and omit the penalty term in the calculation of (3.1), which amounts to non-penalised local linear regression on variables indexed by $\{d : \hat{h}_d \le C_0\}$ using variable-specific bandwidths $\hat{h}_d$'s. We comment briefly on the practical choice of C_0 in Section 7.

Remark 1. The choice of the penalty function $\Lambda_n(h)$ is a trade-off between smoothness of the objective function and identifiability of the active set $\mathcal{A}$. Our theoretical results remain valid under the weaker assumption that Λ_n is strictly decreasing and satisfies

$$\Lambda_n(h)/\lambda_n = \begin{cases} 1 + O(h^4), & h \to 0 \\ o(1), & h \to \infty. \end{cases}$$

Remark 2. The strategy of pushing bandwidths to infinity for inactive variables allows D to be high dimensional, thus facilitating the construction of an oracle estimator. By contrast, if the bandwidths h_d's were non-adaptive to data, as has been considered by Bertin & Lecué (2008), the number of “local points” would have been inversely proportional to the usual “variance term”, of order $n^{-1}\prod_{d=1}^{D} h_d^{-1}$. Consider for brevity the case where $X^{(1)}, \ldots, X^{(D)}$ are i.i.d. with $\mathrm{supp}(X^{(d)}) = [-a, a]$ and that the kernel K is uniform on a bounded support $[-1, 1]$. If the bandwidths h_d's are all non-shrinking with values exceeding a, then the kernel weight $\prod_{d=1}^{D} K(X^{(d)}/h_d)$ reduces to the constant 2^{-D}, which amounts to the classical case of linear regression. On the other hand, if $h_d \to 0$

for each d, then we have a local linear fit prone to the curse of dimensionality, where the number of local points could become extremely small or even zero for a moderately large D. Even with a small D, setting $h_d \to 0$ for an inactive $d \notin \mathcal{A}$ incurs an undesirable efficiency loss by reducing the effective sample size by a factor of h_d. In order to achieve a level of optimality comparable to that of the oracle, one must minimise the adverse effects the inactive variables may have on the estimator (3.1), which in our procedure is accomplished by forcing the inactive bandwidths to infinity with a judiciously specified penalty. The active variables can be readily selected based on the magnitudes of the optimised bandwidths, an approach fundamentally different from that of Bertin & Lecué (2008), who resort to penalisation of linear coefficients for variable selection. This is also the reason why the choice of the power α in (3.1) is immaterial as long as the penalty is appropriately adjusted by C_n and (3.1) yields a consistent estimator of the regression function with appropriate bandwidths.

We conclude this section with a brief comparison between our procedure and existing methods of the local polynomial type for variable selection and regression. In RODEO (Lafferty & Wasserman, 2008), optimality of estimation of regression coefficients is impeded by the absence of penalties, especially when the number of inactive variables is large. Another disadvantage is its inability to detect linear effects. Such disadvantage is also faced by Giordano & Parrella (2016) and Giordano et al. (2020), who extend the RODEO idea to a high-dimensional context. Both papers resort to a scheme of multiple steps to identify all active variables, under very restrictive distributional assumptions on $\boldsymbol{X}$. In contrast, estimation of $\boldsymbol{h}$ and β_d's in our procedure is regularised by C_n and Λ_n simultaneously, such that oracle optimality ensues without omission of active linear effects, all accomplished by a unified optimisation scheme under only mild conditions on the distribution of $\boldsymbol{X}$. Specifically, our optimisation scheme includes a penalty weighted by $C_n(h_d)$, which shrinks β_d towards zero when h_d is sufficiently large, with a result akin to performing local constant regression on the d-th dimension. It can be shown using standard asymptotic arguments that if $\nabla_d f(\boldsymbol{X})$ does not vanish almost surely, then a local constant fit is consistent only when the bandwidths associated with linear variables converge to zero. Thus, under the assumption (A2) to be introduced in Section 5, our method ensures consistent selection of both linear and nonlinear active variables by thresholding the estimated bandwidths $\hat{\boldsymbol{h}}$. Bertin & Lecué (2008) estimate local linear coefficients subject to an L_1 penalty, with bandwidths $\boldsymbol{h}$ fixed at some pre-determined values. Their procedure succeeds in selecting variables that are active locally at a test point $\boldsymbol{x}$, under a modest dimension $D = O(\log n)$. Our alternating optimisation scheme automatically adjusts $\boldsymbol{h}$ to yield a set of kernel weights adaptive to the selected variables, resulting in a consistent estimate of $\mathcal{A}$, the set of globally active variables, under dimensions D much greater than n. The MEKRO (White et al., 2017) subjects inverse bandwidths to a constraint which not only depends on the unknown sparsity level $|\mathcal{A}|$ but also forces an imbalance between bias and variance, leading to a suboptimal convergence rate. Our procedure does not have such problems and, based as it is on a local linear fit, incurs smaller boundary bias than does MEKRO. The above advantages over existing methods render our procedure especially attractive for global variable selection and prediction under high dimensions.

4. Computational algorithm

The objective function in (3.1) is convex in $(\beta_0, \boldsymbol{\beta}^\top)$ and can be readily minimised by standard optimisation algorithms, the choice of which may be made specific to the loss function L. The objective function in (3.2) is non-convex in $\boldsymbol{h}$ and the solution $\hat{\beta}_0^{-\mathcal{S}_k}(\boldsymbol{X}_i; \boldsymbol{h})$ found using (3.1) has no closed form in general. It is hard to apply standard optimisation methods to obtain (3.2). We propose a coordinate descent method associated with step size ς to find $\hat{\boldsymbol{h}}$. In the algorithm that follows, each bandwidth h_d is replaced by a scaled bandwidth $g(h) = \{1 + \log(1 + h^4)\}^{-1}$, which decreases in h with $g(h) = 0$ when $h = \infty$. Write $\boldsymbol{g}(\boldsymbol{h}) = \big(g(h_1), \ldots, g(h_D)\big)^\top$ for the scaled bandwidth vector. In each iteration, we find the step direction in each dimension which reduces the value of the objective function in (3.2) until a stopping criterion is reached. Based on the output $\boldsymbol{g}(\boldsymbol{h})$, we consider the variable d active whenever $g(h_d) \geq g(C_0)$, for some constant $C_0 > 0$. We outline the algorithm below.

Step 1. Initialise step size ς, toleration limit ε, the minimum scaled bandwidth $g_{min} = 0$, the maximum scaled bandwidth g_{max} to be smaller than but close to 1, and $\boldsymbol{h} = \boldsymbol{h}_0$. Calculate the initial scaled bandwidth vector $\boldsymbol{g}(\boldsymbol{h}_0)$.

Step 2. Calculate the initial $\hat{\beta}_0^{-\mathcal{S}_k}(\boldsymbol{X}_i; \boldsymbol{h}_0)$'s by solving (3.1) for each $i = 1, 2, \ldots, n_0$, $k = 1, \ldots, K$ and hence the initial value f_1 of the objective function in (3.2).

Step 3. Set $f_2 = f_1$. Generate a random permutation $(v_1, \ldots, v_D)$ of $(1, \ldots, D)$. For $d = 1, \ldots, D$, repeat *Steps 3.1–3.2* below.

Step 3.1. Set the lower and upper test values of $g(h_{v_d})$ to be $g_{v_d}^- = g_{min} \vee \big(g(h_{v_d}) - \varsigma\big)$ and $g_{v_d}^+ = g_{max} \wedge \big(g(h_{v_d}) + \varsigma\big)$, respectively. Let f_2^- and f_2^+ be the values of the objective function in (3.2), evaluated at $\boldsymbol{h}$ with h_{v_d} replaced by $g^{-1}(g_{v_d}^-)$ and $g^{-1}(g_{v_d}^+)$, respectively.

Step 3.2. Update (h_{v_d}, f_2) to $\big(g^{-1}(g_{v_d}^-), f_2^-\big)$ if $f_2^- < f_2$; otherwise update (h_{v_d}, f_2) to $\big(g^{-1}(g_{v_d}^+), f_2^+\big)$ if $f_2^+ < f_2$.

Step 4. If $(f_1 - f_2)/f_1 < \varepsilon$, output the scaled bandwidth vector $\boldsymbol{g}(\boldsymbol{h})$ and f_2; otherwise set $f_1 = f_2$ and repeat *Step 3*.

Although the greedy algorithm is likely to be trapped by a local solution, the output $\boldsymbol{h}$ provides a reasonably good approximation to $\hat{\boldsymbol{h}}$. Empirical performance of this algorithm is illustrated through a number of numerical studies reported in Section 8. To reduce the chance of being trapped by suboptimal local minima, the algorithm may be repeated with multiple initial guesses. We do not pursue this in our empirical studies, as trial runs of the approach reveal only limited gain in predictive accuracy.

5. Theory

Define, for $s \in \mathscr{Y}$, $t \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^D$,

$$q_1(s,t) = \partial L(s,t)/\partial t, \quad \eta(t|\boldsymbol{x}) = \mathbb{E}[q_1(Y,t)|\boldsymbol{X} = \boldsymbol{x}],$$
$$q_3(t|\boldsymbol{x}) = \partial^2 \eta(t|\boldsymbol{x})/\partial t^2,$$
$$\sigma^2(\boldsymbol{x}) = \mathrm{Var}\big(q_1(Y, m(\boldsymbol{x}))\big|\boldsymbol{X} = \boldsymbol{x}\big) \quad \text{and} \quad v(\boldsymbol{x}) = \partial \eta(t|\boldsymbol{x})/\partial t|_{t=m(\boldsymbol{x})},$$

provided that the derivatives exist.

For $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{c} = (c_1, \ldots, c_D)^\top \in [0,\infty]^D$, define $\mathcal{N}(\boldsymbol{c}) = \{d : c_d = 0\}$ and

$$r(\boldsymbol{x}; \boldsymbol{c}) = \operatorname*{argmin}_{\beta_0 \in \mathbb{R}} \mathbb{E}\Big[\mathbb{E}\big[L(Y,\beta_0)\big|\boldsymbol{X} = \boldsymbol{x} + \boldsymbol{U}^{\boldsymbol{x},\boldsymbol{c}}\big]\Big|\boldsymbol{U}^{\boldsymbol{x},\boldsymbol{c}}_{\mathcal{N}(\boldsymbol{c})} = \boldsymbol{0}\Big], \tag{5.1}$$

where $\boldsymbol{U}^{\boldsymbol{x},\boldsymbol{c}} \in \mathbb{R}^D$ has the density function $\propto f(\boldsymbol{x}+\boldsymbol{u}) \prod_{d \notin \mathcal{N}(\boldsymbol{c})} K(u_d/c_d)$, $\boldsymbol{u} \in \mathbb{R}^D$. We shall show in Lemma 1 that $r(\boldsymbol{x}; \boldsymbol{c})$ is related to the asymptotic limit of $\hat{\beta}_0^{-\emptyset}(\boldsymbol{x}; \boldsymbol{h})$ for a general choice of $\boldsymbol{h}$ which does not necessarily provide for consistency of $\hat{\beta}_0^{-\emptyset}(\boldsymbol{x}; \boldsymbol{h})$. If $c_j = 0$ for all $j \in \mathcal{A}$, that is $\mathcal{A} \subset \mathcal{N}(\boldsymbol{c})$, then $r(\boldsymbol{x}; \boldsymbol{c}) = \operatorname{argmin}_{\beta_0} \mathbb{E}\big[L(Y,\beta_0)\big|\boldsymbol{X}_{\mathcal{A}} = \boldsymbol{x}_{\mathcal{A}}\big] = m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$. In the special case where L is the squared loss, $r(\boldsymbol{x}; \boldsymbol{c})$ reduces to the conditional expectation $\mathbb{E}\big[m_{\mathcal{A}}\big(\boldsymbol{x}_{\mathcal{A}} + \boldsymbol{U}^{\boldsymbol{x},\boldsymbol{c}}_{\mathcal{A}}\big)\big|\boldsymbol{U}^{\boldsymbol{x},\boldsymbol{c}}_{\mathcal{N}(\boldsymbol{c})} = \boldsymbol{0}\big]$.

We make the following assumptions.

(A1) $f(\cdot)$ and $\sigma(\cdot)$ are twice continuously differentiable, $m(\cdot)$ is two times continuously differentiable, and

$$\mathbb{E}\Big[\|\boldsymbol{X}\|_2^3\big\{v(\boldsymbol{X})|\nabla_{d_1,d_2} m(\boldsymbol{X})| + \sigma(\boldsymbol{X})^2 + |q_1(Y, m(\boldsymbol{X}))|^{2+\delta_0} + 1\big\}\Big] < \infty,$$

for some $0 < \delta_0 \leq 1$ and for any $d_1, d_2 \in \mathcal{A}$.

(A2) $\mathbb{P}(\nabla_{d,d} m(\boldsymbol{X}) = 0)\mathbb{P}(\nabla_d f(\boldsymbol{X}) = 0) < 1$ for each $d \in \mathcal{A}$.

(A3) $q_1(Y,t)$, $q_3(t|\boldsymbol{X})$ and $\eta(t|\boldsymbol{X})$ exist almost surely, for any $t \in \mathbb{R}$.

(A4) $\mathbb{E}\Big[\big\{L(Y, m(\boldsymbol{X}) + t) - L(Y, m(\boldsymbol{X})) + t q_1(Y, m(\boldsymbol{X}))\big\}^2\Big|\boldsymbol{X}\Big] = o(t^2)$ almost surely.

(A5) $q_3(\cdot|\boldsymbol{X})$ is bounded on a neighbourhood containing $m(\boldsymbol{X})$ almost surely.

(A6) $q_1(Y, \cdot)$ is Lipschitz continuous in the sense that there exists a sub-exponential $L_0(Y)$ such that $|q_1(Y,t_1) - q_1(Y,t_2)| \leq L_0(Y)|t_1 - t_2|$ for all $t_1, t_2 \in \mathbb{R}$.

(A7) $L(Y,0)$ and $q_1(Y,0)$ are sub-exponential conditional on $\boldsymbol{X}$, with the associated Orlicz norm uniformly bounded almost surely.

(A8) $m_{\mathcal{A}}$ satisfies

$$\sum_{d_1,d_2 \in \mathcal{A}} \sup_{\boldsymbol{u} \in \mathrm{supp}(\boldsymbol{X}_{\mathcal{A}})} |\nabla_{d_1,d_2} m_{\mathcal{A}}(\boldsymbol{u})| \leq \omega,$$

for some constant $\omega \in [1, |\mathcal{A}|^2]$.

Standard for local polynomial regression, (A1) imposes mild regularity conditions on the distribution of $(\boldsymbol{X}, Y)$ and smoothness conditions on m and L. The $(2+\delta_0)$-moment is required here for establishing asymptotic normality. As discussed earlier, (A2) ensures that the optimal bandwidths converge to zero if and only if they are associated with active variables, be they linear or nonlinear, so that the estimated bandwidths $\hat{\boldsymbol{h}}$ help differentiate active from inactive variables. The assumption (A2) is very mild, for it rules out only a very special case where $\nabla_{d,d} m(\boldsymbol{X}) = \nabla_d f(\boldsymbol{X}) = 0$ almost surely for some $d \in \mathcal{A}$, under which the bias of local constant regression no longer increases with h_d for some linear variables $X^{(d)}$. Existence of $\eta(t|\boldsymbol{X})$ under (A3), required too by Fan et al. (1994), allows for applications to robust regression. Existence of $q_3(t|\boldsymbol{X})$ follows from twice-differentiability of $\eta(\cdot|\boldsymbol{X})$, which holds in particular for mean and logistic regressions. For quantile regression, $q_3(t|\boldsymbol{X})$ exists if the conditional distribution of Y given $\boldsymbol{X}$ admits a differentiable density almost surely. The assumption (A4) also appears in Fan et al. (1994) and is required for consistent estimation of the regression function. It ensures smoothness of the expected loss function conditional on $\boldsymbol{X}$, and is satisfied if $q_1(Y, t)$ is uniformly continuous in t, as is the case for the squared loss or a smooth negative loglikelihood. The check loss also satisfies (A3) as q_3 exists under (A2). The mild condition (A5) ensures that $\mathbb{E}[L(Y, t)|\boldsymbol{X}]$ is well approximated by a quadratic Taylor expansion. It holds for a continuous $q_3(\cdot|\boldsymbol{X})$ or for quantile regression under a bounded conditional density derivative in a neighbourhood of $m(\boldsymbol{X})$. The assumption (A6) implies that $\sup_{t_1,t_2\in\mathbb{R},\, 1\le i\le n} |q_1(Y_i, t_1) - q_1(Y_i, t_2)|/|t_1 - t_2| = O_p(\log n)$, and is satisfied by most generalised linear models including normal and logistic regressions but does not hold for quantile regression. The assumption (A7) imposes on the conditional distribution of Y given $\boldsymbol{X}$ a tail condition specific to the loss function. For example, a squared loss requires the conditional distribution to be sub-exponential almost surely. The additional assumption (A8) on the smoothness of the regression function is commonly made in the context of nonparametric regression for deriving the minimax rate (Yang & Tokdar, 2015). The upper bound ω is imposed to exclude highly fluctuating functions.

The following conditions delimit the sparsity level $|\mathcal{A}|$ and the covariate dimension D, both of which may increase with n.

(B1) $|\mathcal{A}|^{5/2} n^{-1/(4+|\mathcal{A}|)} = o(1)$;
(B2) $D = O\big(e^{\zeta_n}\big)$ for some $\zeta_n = o\big(n^{1/(4+|\mathcal{A}|)}\big)$.

Under (B1), $|\mathcal{A}|$ can be as large as $a_0 \log n / \log\log n$ for any constant $a_0 \in (0, 2/5)$. The conditions (B1) and (B2) together allow a high dimension D with $\log D = O\big((\log n)^{a_1}\big)$ for any constant $a_1 \in (0, 2/5)$, so that D may grow faster than any polynomial rate. If $|\mathcal{A}| = O(1)$, then a_1 can be relaxed to any constant in $(0, \infty)$.

For results involving an explicit leading term of the bias of local linear regression, we simplify our presentation by strengthening the condition (A2) to

(A2') $\mathbb{P}(\nabla_{d,d} m(\boldsymbol{X}) = 0) < 1$ for all $d \in \mathcal{A}$.

If (A2) holds but (A2') does not, then our procedure remains consistent in detecting $\mathcal{A}$ but the bias of local linear regression has a complicated leading term depending on $\nabla_{d,d'}m(\boldsymbol{x})$ for $d, d' \in \mathcal{A}$, which prohibits a simple formulation of the optimal bandwidths.

Theorem 1 below establishes consistency of $\hat{\mathcal{A}}$ and the asymptotic order of the estimated bandwidth vector $\hat{\boldsymbol{h}}$ defined in (3.2).

Theorem 1. *Assume conditions (A1)–(A8) and (B1)–(B2). Then,*

(i) $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1.$

Further, if (A2') holds, then

(ii) there exist constants $K_0 > k_0 > 0$ *such that*

$$\mathbb{P}\Big(k_0(n\omega^2)^{-1/(4+|\mathcal{A}|)} \le \min_{d\in\mathcal{A}} \hat{h}_d \le \max_{d\in\mathcal{A}} \hat{h}_d \\ \le K_0(n\omega^2)^{-1/(4+|\mathcal{A}|)}\Big) \to 1;$$

(iii) for a sufficiently small constant $m_0 > 0$,

$$\mathbb{P}\Big(\min_{d\in\mathcal{A}^c}\{\log \hat{h}_d\} \ge m_0\lambda_n\omega^2(\omega^2 n)^{4/(4+|\mathcal{A}|)}\Big) \to 1.$$

Part (i) of Theorem 1 confirms variable selection consistency of our procedure. As can be seen from the proof, such consistency follows essentially from the fact that the estimated bandwidth $\hat{h}_d = o_p(1)$ if and only if $d \in \mathcal{A}$. If we strengthen (A2) to (A2'), then parts (ii) and (iii) assert that with probability converging to one, componentwise $\hat{\boldsymbol{h}}_{\mathcal{A}}$ has the oracle optimal order, whereas all the components of $\hat{\boldsymbol{h}}_{\mathcal{A}^c}$ explode to infinity at a fast rate.

We investigate next the asymptotic properties of the regression function estimator $\hat{m}(\boldsymbol{x}) = \hat{\beta}_0^{-\emptyset}(\boldsymbol{x}; \hat{\boldsymbol{h}})$ derived from (3.1) with $\boldsymbol{h} = \hat{\boldsymbol{h}}$.

Theorem 2. *Under the conditions of Theorem 1, we have, for any* $\boldsymbol{x} \in \mathbb{R}^D$,

(i) $\hat{m}(\boldsymbol{x}) = m(\boldsymbol{x}) + O_p\big(\omega(n\omega^2)^{-2/(4+|\mathcal{A}|)}\big)$;

(ii) $\mathbb{E}\big[L(Y, \hat{m}(\boldsymbol{x}))\big|\boldsymbol{X} = \boldsymbol{x}, \mathscr{D}_n\big] = \mathbb{E}\big[L(Y, m(\boldsymbol{x}))\big|\boldsymbol{X} = \boldsymbol{x}, \mathscr{D}_n\big] + O_p\big(\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)}\big)$, *for* $(\boldsymbol{X}, Y)$ *independent of* $\mathscr{D}_n$;

(iii) if (A2) is strengthened to (A2') and the kernel function K has a bounded support, then

$$n^{1/2}\Big(\prod_{d\in\mathcal{A}} \hat{h}_d\Big)^{1/2}\Big\{\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) - (\mu_{1,2}/2)\sum_{d\in\mathcal{A}} \hat{h}_d^2 \nabla_{d,d} m(\boldsymbol{x})\Big\}$$

converges in distribution to a normal random variable with mean 0 and variance $\mu_{2,0}^{|\mathcal{A}|}\sigma(\boldsymbol{x})^2 f_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})^{-1}\upsilon(\boldsymbol{x})^{-2}$, *where* $f_{\mathcal{A}}$ *denotes the marginal density function of* $\boldsymbol{X}_{\mathcal{A}}$.

Parts (i) and (ii) of Theorem 2 show that the estimation and prediction errors of $\hat{m}(\boldsymbol{x})$ achieve the minimax rates established by Yang & Tokdar (2015), respectively. In particular, if ω is a fixed positive constant, the results give the classical rates found in Ruppert & Wand (1994). Part (iii) establishes asymptotic normality of the error $\hat{m}(\boldsymbol{x})-m(\boldsymbol{x})$, properly centred and scaled, under the stronger assumption (A2'). It implies an oracle property for $\hat{m}(\boldsymbol{x})$, which enjoys a very low extra bias and a minimax convergence rate, as a consequence of Theorem 1(ii) and (iii) and use of a specially constructed penalty function defined in 3.1. In particular, for generalised linear models (2.1), we have $v(\boldsymbol{x}) = b''(m(\boldsymbol{x}))$ and $\sigma(\boldsymbol{x})^2 = \mathrm{Var}(Y|\boldsymbol{X} = \boldsymbol{x})$.

Remark 3. The uniform lower bound on the orders of $\{\log \hat{h}_d : d \in \mathcal{A}^c\}$ given in Theorem 1(iii) holds for the particular penalty $\Lambda_n(h) = \lambda_n \left\{1 + \log\left(1 + h^4\right)\right\}^{-1}$. Other choices of Λ_n satisfying the general conditions described in Remark 1 may lead to different forms of lower bounds on the orders of $\hat{h}_d$, $d \in \mathcal{A}^c$.

Remark 4. We can show that Theorem 1(i) and (iii) still hold even if (A2) is violated. Note that the leading term of the bias of local linear regression depends on the second derivatives, $\{\nabla_{d,d}m(\boldsymbol{X}) : d \in \mathcal{A}\}$, of the regression function. If $\nabla_{d,d}m(\boldsymbol{X})$ vanishes, the order of the bias becomes smaller, dropping from h_d^2 to, for example, $O(h_d^4)$ or $o(n^{-1}\prod_{d\in\mathcal{A}} h_d^{-1})$, the precise expression of which depends on the regression function m and the density function f. Such reduction in bias results in bigger optimised bandwidths $\hat{h}_d$ for $d \in \mathcal{A}$, which still converge to zero simultaneously with high probability in the sense of Theorem 1(ii), albeit at a slower rate.

6. Quantile regression

Nonparametric quantile regression constitutes an important subclass of problems to which we can apply our proposed procedure. For a fixed $\tau \in (0,1)$, τth-quantile regression amounts to setting the loss function $L(Y,a) = L_\tau(Y-a)$, where $L_\tau(z) \triangleq \tau z \mathbf{1}\{z \geq 0\} - (1-\tau)z\mathbf{1}\{z < 0\}$ is known commonly as the check function. The regression function $m(\boldsymbol{X})$ then corresponds to the conditional τth-quantile of Y given $\boldsymbol{X}$. High dimensional L_1-penalised quantile regression for linear models has been studied in van de Geer (2008) and Belloni & Chernozhukov (2011), where both the number of regressors and the size of active set are allowed to grow with sample size n. Lin et al. (2013) extend COSSO (Lin & Zhang, 2006) to additive quantile regression for variable selection under a fixed dimension. The loss functions covered by Allen's (2013) KNIFE procedure include the check loss as a special case. In a related context, Fan et al. (1994) discuss the use of local linear fitting in univariate robust regression. To the best of our knowledge, the problem of variable selection and estimation for high-dimensional nonparametric quantile regression remains unexplored.

Since the check function does not satisfy (A6), which is used for establishing (S.16) in the proof of Lemma 2, the results of Theorems 1 and 2 do not apply immediately to quantile regression. The problem can nevertheless be resolved by either replacing (B1) and (S.17) with the stronger condition $|\mathcal{A}| = O(1)$ and (S.18), respectively, or by using a local constant fit instead of a local linear fit in (3.1). In either case, we can apply standard empirical process techniques to establish Lemmas 1 and 2 and, following which, Theorems 1 and 2. The orders of the estimated bandwidths remain unchanged and, in the case where $\hat{m}(\boldsymbol{x})$ is obtained by a local constant fit, the bias term $(\mu_{1,2}/2)\sum_{d\in\mathcal{A}}\hat{h}_d^2\nabla_{d,d}m(\boldsymbol{x})$ in Theorem 2(iii) should be changed to $\sum_{d\in A}\hat{h}_d^2\big\{\nabla_d(f_{\mathcal{A}}v)(\boldsymbol{x})\nabla_d m(\boldsymbol{x})/(f_{\mathcal{A}}v)(\boldsymbol{x})+(\mu_{1,2}/2)\nabla_{d,d}m(\boldsymbol{x})\big\}$. Indeed, the proof for the case of a local constant fit largely repeats what has been presented in the Appendix and can be derived in the same manner. Note that for τth-quantile regression, $v(\boldsymbol{x})$ reduces to the conditional density of Y at $m(\boldsymbol{x})$, given $\boldsymbol{X}=\boldsymbol{x}$, and $\sigma(\boldsymbol{x})^2$ reduces to the constant $\tau(1-\tau)$.

7. Choice of tuning parameters

Our proposed method depends on several parameters, which include c in the penalty weight $C_n(h_d)$, λ_n in the bandwidth penalty $\Lambda_n(\cdot)$, as well as the nonlinear variable selection threshold C_0.

Recall that $C_n(h_d)$ penalises the local linear coefficient β_d with a weight which grows with the corresponding bandwidth h_d at a rate controlled by the parameter c. To balance the smoothness of the objective function in (3.1) against the growing rate of $C_n(h_d)$, c is fixed to be 10 in our empirical studies.

We have established in Theorem 1 that $h_d\to 0$ for $d\in\mathcal{A}$ and $h_d\to\infty$ for $d\notin\mathcal{A}$ with high probability, suggesting that C_0 can be any fixed positive constant. Our numerical results show that $\hat{\mathcal{A}}$ remains stable over a wide range of C_0. We fix C_0 to be 1 in our empirical studies.

Compared to c and C_0, choice of λ_n makes a relatively large impact on selection and prediction outcomes. We suggest fixing λ_n by cross validation, which is found to work satisfactorily in practice.

We present in Section 8 some numerical evidence in support of the above recommendations under the setting of Example M1.

As our theoretical results hold for any $\alpha\geq 1$, the choice of α is made out of concern for computational efficiency of the procedure. It is often computationally more efficient to solve (3.1) based on the choice $\alpha=2$, particularly for the case of mean regression where a solution exists in closed form. However, one may also consider other choices to achieve specific goals; for example, setting $\alpha=1$ helps yield sparse coefficients. In our numerical studies we set $\alpha=2$.

8. Empirical studies

We conduct empirical studies to illustrate our proposed penalised local linear regression procedure, abbreviated henceforth as PLLR, for mean, quantile and logistic regressions using both simulated and real data. Its performance is measured by variable selection accuracy and predictive power. Cross validation is performed to choose the best step size and λ_n over a grid of trial values. We include in the studies a local constant variant on the procedure, with $\beta_1=\cdots=\beta_D=0$ and abbreviated as PLCR, whereby (3.1) reduces to $\hat{\beta}_0^{-\mathcal{S}}(\boldsymbol{x};\boldsymbol{h})=\operatorname{argmin}_{\beta_0\in\mathbb{R}}\sum_{i\notin\mathcal{S}}K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})L\big(Y_i,\beta_0\big)$. Three loss functions are considered, corresponding respectively to mean, quantile and logistic regressions, for which (3.1) is solved by the R packages glmnet and rqPen. We fix $c = 10$ for the penalty weight $C_n(h)$ and $\alpha = 1$ in (3.1). The selected set $\hat{\mathcal{A}}$ is set to be $\{d : \hat{h}_d \leq 1\}$.

For the other methods we set the regularising parameters by either cross validation or AICc available in their corresponding codes or packages. Here AICc is a modified version of AIC designed for small samples and is defined to be $\log(\mathsf{r}) + \{n + \mathrm{tr}(\mathscr{S})\}/\{n - \mathrm{tr}(\mathscr{S}) + 2\}$, where r and $\mathscr{S}$ denote the mean squared residuals and the Nadaraya–Watson linear smoother, respectively.

Throughout the studies, all covariates have been standardised prior to the fitting of each model. The high-dimensional setting $D > n$ is considered in some of the simulation and real data examples.

8.1. Mean regression

We consider three simulation examples and one real data example, in which PLLR and PLCR are compared with MEKRO (White et al., 2017), LASSO, KNIFE (Allen, 2013), COSSO (Lin & Zhang, 2006), ACOSSO (Storlie et al., 2011) and GRID (Giordano et al., 2020). Among the eight methods, PLLR, PLCR, GRID and MEKRO are of the local polynomial type, LASSO is tailored for linear models, while KNIFE, COSSO and ACOSSO model the regression function by RKHS. An additive structure is further assumed by COSSO and ACOSSO. All simulation results are obtained from 100 Monte Carlo replications.

Example M1. Consider a homoscedastic model where

$$X^{(d)} = (U_d + tU)/(1 + t), \; d = 1, \ldots, 200,$$

Y is conditionally Gaussian with mean $20(X^{(20)} - 0.5)\cos(0.7X^{(1)} + 1 - \pi)$ and variance 0.25 given $\boldsymbol{X}$, and $U, U_1, \ldots, U_{100}$ are independently and uniformly distributed over $(0, 1)$. The parameter t is fixed at either 0 or 0.5, so that the pairwise correlations (ρ) between covariates are 0 or 0.2, respectively. The sample size n is set to be 150.

The numbers of cases of correct selection ($\hat{\mathcal{A}} = \mathcal{A}$), false positives only ($\hat{\mathcal{A}} \supsetneq \mathcal{A}$), false negatives only ($\hat{\mathcal{A}} \subsetneq \mathcal{A}$) and both false positives and negatives are shown in Table 1. Results for the COSSO and adaptive COSSO are unavailable as their R codes fail to run in the present high-dimensional setting. As seen from Table 1, PLLR is most accurate in variable selection among all methods and never experiences false negatives. Its close relative PLCR is slightly worse when $\rho = 0$ and much worse when $\rho = 0.2$. The other methods perform poorly in variable selection.

We also calculate the average prediction error, defined to be $n_1^{-1} \sum_{i=1}^{n_1} \left|\hat{m}(\boldsymbol{Z}_i) - m(\boldsymbol{Z}_i)\right|^j$, of each method, over $n_1 = 100$ test points $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{n_1}$ generated independently from the same distribution as $\boldsymbol{X}$. Table 1 reports the mean squared error (MSE) and the mean absolute error (MAE), obtained by setting $j = 2$ and 1, respectively. Serving as a benchmark, the oracle method

Table 1

Example M1 (mean regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

		variable selection				prediction	
ρ	Method	CS	FP	FN	FPN	MSE	MAE
0	PLCR	67	33	0	0	1.56	0.875
	PLLR	**88**	12	0	0	**0.441**	**0.485**
	MEKRO	2	14	66	18	1.82	0.969
	GRID	0	44	7	49	1.99	0.985
	KNIFE	0	58	10	32	1.54	0.837
	LASSO	0	0	85	5	1.61	0.928
	Oracle	–	–	–	–	0.342	0.402
0.2	PLCR	0	78	12	10	0.698	0.584
	PLLR	**85**	15	0	0	**0.242**	**0.348**
	MEKRO	5	0	60	35	0.712	0.635
	GRID	0	25	12	63	0.624	0.599
	KNIFE	0	44	0	56	0.581	0.498
	LASSO	0	0	81	19	0.533	0.500
	Oracle	–	–	–	–	0.218	0.315

Table 2

Example M1 (mean regression, $\rho = 0$) — effects of varying C_0 and c on prediction and variable selection, measured by absolute percentage change (abs. pct. diff) in $\hat{m}\big(\boldsymbol{Z}_i\big)$ and rate of correct selection with $\hat{\mathcal{A}} = \mathcal{A}$, averaged over 100 replications.

C_0	0.5	0.75	1	2	5
avg. pct. diff.	10%	0%	0%	2%	14%
Rate of correct selections	78%	90%	88%	87%	80%

c	1	5	10	20	50
avg. pct. diff.	26%	2%	0%	1%	2%
Rate of correct selections	70%	90%	90%	90%	88%

estimates $m(\boldsymbol{x})$ by $\hat{\beta}_0^{-\emptyset}(\boldsymbol{x};\hat{\boldsymbol{h}})$, which is derived from (3.1) under the oracle constraint $\|\boldsymbol{\beta}_{\mathcal{A}^c}\|_1 = 0$, with $\hat{\boldsymbol{h}}$ set by cross validation. Being closest to the oracle, PLLR outperforms the other four methods in both MSE and MAE. In general, the prediction errors for $\rho = 0.2$ are smaller than those for $\rho = 0$, since inactive variables help explain the response partially.

We have conducted an additional simulation exercise to examine the effects of varying the bandwidth threshold C_0 and the parameter c on our method under the setting of this example. Two indicators are reported: (1) absolute percentage change in $\hat{m}\big(\boldsymbol{Z}_i\big)$, calculated by averaging over 100 test points, and (2) rate of correct selection with $\hat{\mathcal{A}} = \mathcal{A}$, both averaged over 100 replications. Table 2 shows that the selection and prediction performances of our method remain quite stable when C_0 or c deviates moderately from the values $C_0 = 1$ or $c = 10$ recommended in Section 7, while the other tuning parameters remain fixed at their recommended values. We also plot in Figure 1 the solution paths of the optimised bandwidths against λ_n. The paths show a large discrepancy in magnitude between bandwidths of the selected and unselected variables near the cross-validated value of λ_n, thus confirming effectiveness of cross validation in fixing λ_n.

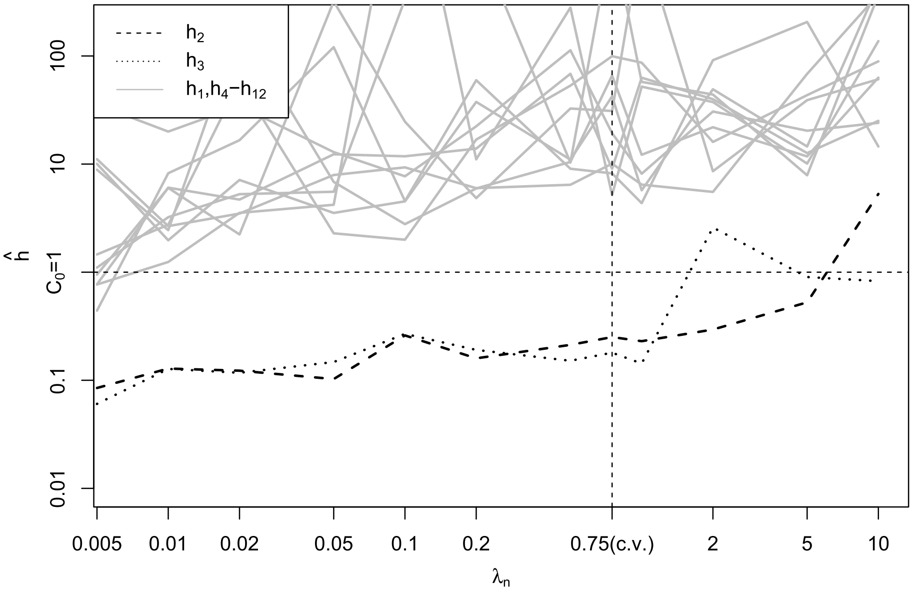

Fig 1. *Example M1 (mean regression, $\rho = 0$) — plot of $\hat{h}_d$ against λ_n, for $d = 1, \ldots, 11$ and 20. The horizontal and vertical dashed lines are drawn at the recommended value of C_0 and the cross-validated value of λ_n, respectively.*

Example M2. Consider the same model as in Example M1, with (D, n) changed to $(500, 300)$. The results shown in Table 3 suggest that the overall selection and estimation performances deteriorate for all methods. However, our proposed method remains the best among all methods and has a satisfactory selection accuracy.

Example M3. Consider a heteroscedastic model with $\boldsymbol{X}$ distributed as in Example M1 and Y normally distributed with mean $20(X^{(20)}-0.5)\cos(0.7X^{(1)}+1-\pi)$ and variance $5\sin^2(2X^{(9)})\sin^2(2X^{(5)})$ conditional on $\boldsymbol{X}$. We set $n = 400$ and $D = 30$. The results, shown in Table 4, suggest that MEKRO and PLLR have the highest selection accuracy for the case $\rho = 0$, whereas PLLR is considerably more accurate than the other methods when $\rho = 0.2$. As the model is neither linear nor additive, LASSO, COSSO and ACOSSO all fail to yield satisfactory selections. Among methods which do not require specific model structures, KNIFE is relatively poor in selection accuracy.

Prediction errors reported in Table 4 show that PLLR has the smallest error, followed closely by MEKRO. The other methods are notably less accurate.

Example M4. Consider an additive heteroscedastic model, where Y is normally distributed with mean $4\sin(3X^{(10)}) - 4\cos(5X^{(5)} - 1.5) + 5(X^{(14)})^2$ and variance $2.5|\cos(3X^{(9)} - 1.5)|$ conditional on $\boldsymbol{X}$, and $\boldsymbol{X}$ follows the same distribution as in Example M1. We set $n = 300$ and $D = 30$. Results on selection accuracy and prediction error are given in Table 5. In terms of selection ac-

Table 3

Example M2 (mean regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

		variable selection				prediction	
ρ	Method	CS	FP	FN	FPN	MSE	MAE
0	PLCR	65	35	0	0	1.41	0.761
	PLLR	**81**	19	0	0	**0.402**	**0.450**
	MEKRO	5	18	60	17	1.77	1.13
	GRID	0	40	8	52	2.08	1.12
	KNIFE	0	42	17	41	1.61	0.908
	LASSO	0	0	91	9	1.51	0.988
	Oracle	–	–	–	–	0.301	0.314
0.2	PLCR	12	78	5	5	0.678	0.588
	PLLR	**79**	21	0	0	**0.248**	**0.351**
	MEKRO	10	6	51	23	0.892	0.687
	GRID	0	28	15	57	0.694	0.608
	KNIFE	0	25	0	75	0.591	0.497
	LASSO	0	0	77	23	0.510	0.518
	Oracle	–	–	–	–	0.187	0.282

Table 4

Example M3 (mean regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

		variable selection				prediction	
ρ	Method	CS	FP	FN	FPN	MSE	MAE
0	PLCR	49	51	0	0	1.58	0.912
	PLLR	67	33	0	0	**0.416**	**0.453**
	MEKRO	**70**	30	0	0	0.610	0.583
	GRID	44	33	3	20	0.824	0.617
	KNIFE	11	16	16	57	1.18	0.847
	COSSO	0	2	70	28	1.39	0.898
	ACOSSO	0	11	29	60	1.44	0.919
	LASSO	0	0	91	9	2.13	1.06
	Oracle	–	–	–	–	0.222	0.336
0.2	PLCR	5	63	16	16	0.572	0.582
	PLLR	**38**	35	17	10	**0.402**	**0.458**
	MEKRO	5	2	80	13	0.432	0.484
	GRID	0	12	35	53	0.467	0.465
	KNIFE	1	5	15	19	0.734	0.622
	COSSO	1	1	75	23	0.445	0.550
	ACOSSO	0	11	26	63	0.484	0.520
	LASSO	0	0	48	52	1.13	0.770
	Oracle	–	–	–	–	0.05	0.329

curacy, PLLR is comparable with COSSO and ACOSSO, methods tailored for additive models. Results of KNIFE, GRID, MEKRO and LASSO are notable for their large numbers of cases plagued by false negatives. For prediction, although inferior to COSSO and ACOSSO, PLLR is on a par with the oracle and more accurate than MEKRO and KNIFE.

Table 5

Example M4 (mean regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

		variable selection				prediction	
ρ	Method	CS	FP	FN	FPN	MSE	MAE
0	PLCR	20	73	4	3	4.05	1.21
	PLLR	62	38	0	0	2.41	1.25
	MEKRO	0	0	100	0	2.89	1.38
	GRID	2	33	11	54	2.99	1.40
	KNIFE	0	0	71	29	3.26	1.45
	COSSO	70	29	1	0	0.560	0.594
	ACOSSO	**68**	21	11	0	**0.351**	**0.431**
	LASSO	3	2	90	5	4.63	1.77
	Oracle	–	–	–	–	1.96	1.18
0.2	PLCR	0	93	0	7	4.59	1.98
	PLLR	**38**	55	3	4	1.48	0.928
	MEKRO	2	1	95	2	2.54	1.28
	GRID	0	7	46	47	2.60	1.31
	KNIFE	0	3	29	8	2.14	1.13
	COSSO	26	46	22	16	0.837	0.677
	ACOSSO	27	43	17	13	**0.549**	**0.542**
	LASSO	1	2	89	8	2.75	1.34
	Oracle	–	–	–	–	1.04	0.805

Example M5. (Real data). The ozone data, available in the R library, have been studied by Lin & Zhang (2006) and Allen (2013). It consists of 330 observations of daily ozone concentration in Los Angeles with 8 predictors. For evaluation of selection performance, we generate 10 artificial predictors independently from the uniform $(0, 1)$ distribution, thus increasing the covariate dimension to $D = 18$. The dataset is partitioned randomly into a training set of 250 observations and a test set $\{(\boldsymbol{X}_{\pi_i}, Y_{\pi_i}) : i = 1, \ldots, n_1\}$ of $n_1 = 80$ observations. Prediction error is evaluated by the mean squared error $n_1^{-1} \sum_{i=1}^{n_1} \{Y_{\pi_i} - \hat{m}(\boldsymbol{X}_{\pi_i})\}^2$. The results, reported in Table 6, are obtained by averaging over 100 random partitions. The method NPLLR refers to conventional, non-penalised, local linear regression on the entire set of 8 genuine covariates without selection. Among the eight methods under study PLLR gives the smallest prediction error. All methods except PLCR are effective in screening out artificial variables.

Example M6. (Real data). The tecator dataset, available at `https://lib.stat.cmu.edu/datasets/`, has been studied by Lin & Zhang (2006) and Storlie et al. (2011). It consists of 240 observations of fat contents with 100 channel spectrum of absorbances. The absorbance is minus the common logarithm of the transmittance measured by the spectrometer. Again, we generate 100 artificial predictors independently from the uniform $(0, 1)$ distribution, thus increasing the covariate dimension to $D = 200$. The traning and testing datasets contain $n = 180$ and 60 samples respectively. The results, reported in Table 7, show that PLLR gives the lowest prediction error among all methods.

Table 6

Example M5 (ozone data, mean regression) — numbers of selected artificial variables and selected variables, and prediction squared error, averaged over 100 random partitions of dataset.

Method	Number of selected artificial variables	variables	Prediction error
PLCR	2.14	9.80	22.7
PLLR	0.89	7.15	**17.1**
MEKRO	0	3.07	20.9
GRID	0.88	3.79	18.9
KNIFE	1.18	4.55	19.0
COSSO	0.08	4.17	19.3
ACOSSO	0.15	3.88	19.2
LASSO	0	4.44	22.4
NPLLR	0	8	26.2

Table 7

Example M6 (tecator data, mean regression) — numbers of selected artificial variables and selected variables, and prediction squared error, averaged over 100 random partitions of dataset.

Method	Number of selected artificial variables	variables	Prediction error
PLCR	2.55	8.73	14.0
PLLR	0.77	6.45	**10.1**
MEKRO	1.21	4.05	12.1
GRID	0.99	6.43	11.9
KNIFE	2.76	8.12	12.5
LASSO	0.42	21.1	12.7
NPLLR	0	200	40.3

8.2. Quantile regression

We compare the same methods as those studied in Section 8.1. Extension of COSSO and ACOSSO to quantile regression is discussed by Lin et al. (2013). The LASSO is taken to be Li & Zhu's (2008) L_1-norm quantile regression. The squared loss is replaced by the check loss in these methods. We perform quantile regression at orders $\tau = 0.2$ and 0.5 in the empirical studies.

Example Q1. Consider the same model and parameter settings as in Example M1. The results are shown in Table 8. Under all the four combinations of (ρ, τ), PLLR outperforms the other three methods (PLCR, KNIFE, LASSO) significantly in both variable selection and prediction.

Example Q2. Consider the same model and parameter settings as in Example M2. The results are shown in Table 9. The selection performances for all methods deteriotiate as compared to Example M1 but PLLR still outperforms the other methods.

Example Q3. Consider the same model and parameter settings as in Example M3. Here heteroscedasticity leads to different active sets $\mathcal{A}$ for different quantile orders τ. In particular, we have $\mathcal{A} = \{1, 5, 9, 20\}$ and $\{1, 20\}$ for $\tau = 0.2$ and

Table 8

Example Q1 (τth-quantile regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

			variable selection				prediction	
ρ	τ	Method	CS	FP	FN	FPN	MSE	MAE
0	0.2	PLCR	42	44	9	5	1.71	0.942
		PLLR	**82**	4	2	12	**0.481**	**0.447**
		KNIFE	22	29	0	49	1.05	0.752
		LASSO	0	24	12	64	1.87	1.01
		Oracle	–	–	–	–	0.057	0.171
	0.5	PLCR	33	62	5	0	1.16	0.723
		PLLR	**87**	0	0	13	**0.444**	**0.389**
		KNIFE	42	40	0	18	0.612	0.584
		LASSO	0	18	19	63	1.42	0.908
		Oracle	–	–	–	–	0.098	0.463
0.2	0.2	PLCR	11	44	18	27	0.95	0.687
		PLLR	**51**	14	11	24	**0.312**	**0.351**
		KNIFE	18	7	61	14	0.632	0.564
		LASSO	0	16	7	77	0.609	0.538
		Oracle	–	–	–	–	0.061	0.162
	0.5	PLCR	4	37	19	40	0.977	0.681
		PLLR	**63**	10	9	18	**0.243**	**0.355**
		KNIFE	22	11	24	43	0.382	0.491
		LASSO	2	15	9	74	0.532	0.513
		Oracle	–	–	–	–	0.050	0.157

Table 9

Example Q2 (τth-quantile regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

			variable selection				prediction	
ρ	τ	Method	CS	FP	FN	FPN	MSE	MAE
0	0.2	PLCR	28	48	12	12	1.84	0.958
		PLLR	**72**	17	6	5	**0.499**	**0.480**
		KNIFE	11	48	0	41	1.33	0.992
		LASSO	0	48	12	40	1.88	1.12
		Oracle	–	–	–	–	0.051	0.151
	0.5	PLCR	28	71	1	0	1.36	0.842
		PLLR	**85**	0	0	15	**0.429**	**0.391**
		KNIFE	28	57	0	15	0.642	0.535
		LASSO	0	17	14	69	1.53	0.958
		Oracle	–	–	–	–	0.091	0.433
0.2	0.2	PLCR	5	62	23	10	1.21	0.727
		PLLR	**44**	16	17	23	**0.487**	**0.430**
		KNIFE	10	8	55	37	0.752	0.590
		LASSO	0	10	10	80	0.801	0.659
		Oracle	–	–	–	–	0.048	0.132
	0.5	PLCR	0	42	25	33	1.38	0.776
		PLLR	**61**	21	12	6	**0.263**	**0.380**
		KNIFE	12	25	5	58	0.477	0.502
		LASSO	0	11	9	80	0.610	0.548
		Oracle	–	–	–	–	0.042	0.133

Table 10

Example Q3 (τth-quantile regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

			variable selection				prediction	
ρ	τ	Method	CS	FP	FN	FPN	MSE	MAE
0	0.2	PLCR	34	41	22	3	2.32	1.21
		PLLR	61	27	8	4	1.07	**0.674**
		KNIFE	**63**	10	19	8	**1.01**	0.771
		COSSO	18	27	44	11	2.53	1.23
		ACOSSO	1	36	24	39	2.50	1.24
		LASSO	1	33	2	64	2.60	1.30
		Oracle	–	–	–	–	0.774	0.549
	0.5	PLCR	78	22	0	0	1.03	0.732
		PLLR	**84**	14	1	1	**0.156**	**0.263**
		KNIFE	70	14	10	6	0.377	0.441
		COSSO	2	37	7	54	1.57	0.955
		ACOSSO	0	23	1	76	1.42	0.904
		LASSO	1	10	11	78	1.38	0.888
		Oracle	–	–	–	–	0.0233	0.103
0.2	0.2	PLCR	0	1	33	66	1.63	0.984
		PLLR	25	44	25	6	0.853	0.705
		KNIFE	**30**	28	35	7	**0.812**	**0.671**
		COSSO	2	9	56	33	0.913	0.743
		ACOSSO	3	22	21	54	0.908	0.747
		LASSO	0	30	4	66	0.995	0.786
		Oracle	–	–	–	–	0.523	0.487
	0.5	PLCR	20	6	59	15	0.346	0.316
		PLLR	**78**	21	1	0	**0.132**	**0.250**
		KNIFE	43	29	18	10	0.267	0.363
		COSSO	8	47	3	42	0.506	0.538
		ACOSSO	2	40	0	58	0.442	0.496
		LASSO	1	31	9	59	0.450	0.492
		Oracle	–	–	–	–	0.0585	0.161

0.5, respectively. Table 10 summarises the results. In general PLLR and KNIFE stand out as the two best performers, with PLLR noticeably better than KNIFE when $\tau = 0.5$ but slightly inferior to KNIFE when $\tau = 0.2$.

Example Q4. The model and parameter settings follow those in Example M4. The results are shown in Table 11. As expected of methods tailored for additive models, both COSSO and ACOSSO perform well in variable selection and prediction. Among methods which do not assume additivity, PLLR appears the best, followed next by KNIFE.

Example Q5. (Real data). The same ozone dataset and partitions as considered in Example M5 are used again here. The prediction error is measured by the mean quantile loss (MQL) $n_1^{-1}\sum_{i=1}^{n_1} L_\tau\big(Y_{\pi_i} - \hat{m}(\boldsymbol{X}_{\pi_i})\big)$. For the median case $\tau = 0.5$, MSE is also calculated as an alternative. Table 12 summarises the results, which show that PLLR selects the fewest variables on average and is the

TABLE 11

Example Q4 (τth-quantile regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE) and mean absolute prediction error (MAE).

			variable selection				prediction	
ρ	τ	Method	CS	FP	FN	FPN	MSE	MAE
0	0.2	PLCR	0	5	49	46	7.57	2.27
		PLLR	11	19	51	19	2.59	1.27
		KNIFE	3	0	81	16	4.15	1.61
		COSSO	**38**	32	27	3	**0.486**	**0.542**
		ACOSSO	15	54	19	12	0.579	0.597
		LASSO	1	20	2	77	5.48	1.84
		Oracle	–	–	–	–	1.35	0.901
	0.5	PLCR	14	44	27	15	4.35	1.70
		PLLR	60	36	3	1	2.19	1.16
		KNIFE	50	5	40	5	2.71	1.31
		COSSO	**89**	11	0	0	**0.143**	**0.291**
		ACOSSO	54	46	0	0	0.189	0.332
		LASSO	0	40	7	53	4.31	1.68
		Oracle	–	–	–	–	0.923	0.767
0.2	0.2	PLCR	0	6	27	67	3.63	1.50
		PLLR	3	20	27	50	1.88	1.02
		KNIFE	1	2	77	20	2.49	1.22
		COSSO	**8**	24	35	33	**0.909**	**0.723**
		ACOSSO	2	25	19	54	1.13	0.808
		LASSO	0	12	1	87	2.79	1.28
		Oracle	–	–	–	–	1.08	0.776
	0.5	PLCR	0	61	6	15	2.89	1.32
		PLLR	31	46	16	7	1.28	0.840
		KNIFE	41	0	53	6	1.97	1.09
		COSSO	**76**	22	2	0	**0.330**	**0.423**
		ACOSSO	31	65	3	1	0.479	0.516
		LASSO	1	35	3	62	2.35	1.18
		Oracle	–	–	–	–	0.767	0.665

most effective in screening out artificial variables. All methods except NPLLR are comparable in terms of predictive accuracy.

Example Q6. (Real data). The same tecator dataset and partitions as considered in Example M6 are used again here. Table 13 summarises the results, which show that PLLR has the lowest prediction error.

8.3. Logistic regression

For logistic regression we compare the methods PLLR, KNIFE, SKDA and LASSO. For moderately large $D \approx 30$, the R code for COSSO produces error messages reporting singular designs so the results are omitted. The LASSO is taken to be L_1-penalised linear logistic regression (Park & Hastie, 2007). Introduced by Stefanski et al. (2014), SKDA can be viewed as the classification version of MEKRO. It maximises the likelihood function under constraints on bandwidths, allowing for a prior on the two classes when estimating the con-

Table 12

Example Q5 (ozone data, τth-quantile regression) — numbers of selected artificial variables and selected variables, mean quantile loss (MQL) and mean squared error (MSE), averaged over 100 random partitions of dataset.

		Number of selected		Prediction error	
τ	Method	artificial variables	variables	MQL	MSE
0.2	PLCR	2	9.14	2.31	–
	PLLR	0.56	3.88	2.22	–
	KNIFE	2.53	7.28	**2.18**	–
	COSSO	1.14	4.77	2.27	–
	ACOSSO	1.76	5.44	2.21	–
	LASSO	4.35	8.63	2.28	–
	NPLLR	0	8	5.68	–
0.5	PLCR	1.89	8.74	3.94	21.3
	PLLR	0.34	4	3.43	19.3
	KNIFE	0.783	5.41	**3.36**	**19.0**
	COSSO	0.415	5.55	3.36	19.2
	ACOSSO	0.915	5.2	3.41	19.6
	LASSO	2.81	7.19	3.65	22.7
	NPLLR	0	8	6.13	55.6

Table 13

Example Q6 (tecator data, τth-quantile regression) — numbers of selected artificial variables and selected variables, mean quantile loss (MQL) and mean squared error (MSE), averaged over 100 random partitions of dataset.

		Number of selected		Prediction error	
τ	Method	artificial variables	variables	MQL	MSE
0.2	PLCR	1.74	9.14	2.02	–
	PLLR	0.87	4.11	**1.84**	–
	KNIFE	2.01	7.28	1.95	–
	LASSO	9.71	18.6	1.87	–
0.5	PLCR	2.02	6.43	3.05	16.3
	PLLR	0.91	4.82	**2.17**	**12.5**
	KNIFE	0.58	6.43	2.44	14.6
	LASSO	6.80	14.25	2.54	14.7

ditional probability. Use of a uniform prior leads to maximum likelihood estimation, abbreviated hereafter as SKDA-mle. Alternatively, one may consider using the Bayes classifier based on data-driven prior weights, leading to the SKDA-bayes approach.

Example L1. Consider a logistic model where, conditional on $\boldsymbol{X}$, Y follows a Bernoulli $\big(p(\boldsymbol{X})\big)$ distribution with $p(\boldsymbol{X})$ given by

$$\begin{aligned}\log\left\{\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right\} &= m(\boldsymbol{X})\\ &= 100(X^{(20)}-0.5)\cos(0.7X^{(9)}+1-\pi)+30(X^{(14)}-0.5).\end{aligned}$$

We set $n=150$, $D=50$ and generate $X^{(1)},\ldots,X^{(D)}$ independently from the uniform $(0,1)$ distribution. Writing $\hat{p}(\boldsymbol{x})=e^{\hat{m}(\boldsymbol{x})}/\big(1+e^{\hat{m}(\boldsymbol{x})}\big)$ for the predicted value of $\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$, we evaluate the prediction error by four measures, namely the MSE $n_1^{-1}\sum_{i=1}^{n_1}\big|\hat{p}(\boldsymbol{Z}_i)-p(\boldsymbol{Z}_i)\big|^2$, the MAE $n_1^{-1}\sum_{i=1}^{n_1}\big|\hat{p}(\boldsymbol{Z}_i)-p(\boldsymbol{Z}_i)\big|$, the misclassification rate (MR) $n_1^{-1}\sum_{i=1}^{n_1}\big|\mathbf{1}\{\hat{p}(\boldsymbol{Z}_i)>0.5\}-\tilde{Y}_i\big|$ and the com-

Table 14

Example L1 (logistic regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE), mean absolute prediction error (MAE), misclassification rate (MR) and comparative Kullback-Leibler distance (CKL).

		variable selection				prediction			
ρ	Method	CS	FP	FN	FPN	MSE	MAE	MR	CKL
0	PLCR	31	2	58	9	0.0692	0.182	0.111	0.345
	PLLR	**86**	13	1	0	**0.0377**	**0.106**	**0.0919**	**0.222**
	KNIFE	46	0	32	22	0.0561	0.172	0.103	0.292
	SKDA-bayes	34	0	65	1	0.0683	0.180	0.134	0.314
	SKDA-mle	34	0	65	1	0.0661	0.180	0.129	0.307
	LASSO	2	3	48	47	0.0845	0.236	0.115	0.372
	Oracle	–	–	–	–	0.0405	0.109	0.0960	0.224
0.2	PLCR	1	0	97	2	0.0772	0.223	0.164	0.385
	PLLR	**62**	22	13	3	**0.0381**	**0.105**	**0.112**	**0.273**
	KNIFE	7	0	72	21	0.0522	0.169	0.119	0.349
	SKDA-bayes	0	0	88	12	0.0738	0.214	0.160	0.383
	SKDA-mle	0	0	87	13	0.0720	0.206	0.153	0.365
	LASSO	1	5	40	54	0.0655	0.206	0.139	0.365
	Oracle	–	–	–	–	0.0349	0.103	0.104	0.212

parative Kullback-Leibler (CKL) distance $n_1^{-1}\sum_{i=1}^{n_1}\big\{-p(\boldsymbol{Z}_i)\log\hat{p}(\boldsymbol{Z}_i)-(1-p(\boldsymbol{Z}_i))\log(1-\hat{p}(\boldsymbol{Z}_i))\big\}$, where $\tilde{Y}_i$ is Bernoulli $\big(p(\boldsymbol{Z}_i)\big)$ distributed conditional on $\boldsymbol{Z}_i$, $i=1,\dots,n_1$. Results are reported in Table 14. We see that PLLR considerably outperforms the other methods in variable selection. In particular, PLLR succeeds in detecting the active variable $X^{(9)}$, exclusion of which has contributed to false negatives in the other methods. In terms of prediction, PLLR again outperforms the other methods by all four measures, and is comparable to the oracle.

Example L2. Consider the same model as in Example L1 except that we set $(n,D)=(300,500)$. The results reported in Table 15 show that PLLR again outperforms the other methods in terms of selection and prediction performance.

Example L3. (Real data). We consider the following three experimental settings, based on datasets available at the UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/`:

(i) Wisconsin diagnostic breast cancer data —
The dataset contains $n=569$ observations, among which 212 are malignant and 357 are benign. The orignal 30 predictors and 20 artificial uniform $(0,1)$ variables together make up a covariate dimension $D=50$. Results are averaged over 100 random partitions of observations into 300 training and 269 test points.

(ii) Sonar data —
The dataset consists of $n=208$ observations and 60 predictors ranging from 0 to 1, indicating the energy within a frequency band over time. Inclusion of 30 artificial uniform $(0,1)$ variables yields a covariate dimension $D=90$. Each observation is classified as 'mineral' (111 in total) or 'rock' (97 in total). Results are averaged over 100 random partitions of

Table 15

Example L2 (logistic regression) — numbers of cases of correct selection (CS), false positives only (FP), false negatives only (FN), both false positives and negatives (FPN), mean squared prediction error (MSE), mean absolute prediction error (MAE), misclassification rate (MR) and comparative Kullback-Leibler distance (CKL).

		variable selection				prediction			
ρ	Method	CS	FP	FN	FPN	MSE	MAE	MR	CKL
0	PLCR	25	17	55	3	0.0790	0.231	0.141	0.478
	PLLR	**81**	8	5	6	**0.0428**	**0.131**	**0.129**	**0.251**
	KNIFE	41	0	41	12	0.0587	0.200	0.141	0.305
	SKDA-bayes	31	0	68	1	0.0713	0.195	0.148	0.328
	SKDA-mle	31	0	68	1	0.0720	0.199	0.157	0.325
	LASSO	0	0	60	40	0.0977	0.258	0.141	0.407
	Oracle	–	–	–	–	0.0355	0.089	0.083	0.210
0.2	PLCR	11	0	87	2	0.0881	0.238	0.174	0.391
	PLLR	**54**	28	10	8	**0.0450**	**0.126**	**0.135**	**0.291**
	KNIFE	0	0	75	25	0.0624	0.187	0.135	0.370
	SKDA-bayes	0	0	84	16	0.0787	0.280	0.174	0.392
	SKDA-mle	0	0	85	15	0.0790	0.269	0.161	0.381
	LASSO	0	10	41	49	0.0710	0.251	0.142	0.410
	Oracle	–	–	–	–	0.0312	0.094	0.091	0.187

observations into 150 training and 58 test points.

(iii) Ionosphere data —

The dataset contains $n = 351$ observations on 35 covariates plus 20 artificial uniform $(0, 1)$ variables. The instances are classified into ‘good’ (225 in total) or ‘bad’ (126 in total). We remove the first two covariates due to lack of heterogeneity. Results are averaged over 100 random partitions of observations into 200 training and 151 test points.

(iv) Musk data —

The dataset contains $n = 476$ observations on 176 covariates plus 150 artificial uniform $(0, 1)$ variables, resulting in a covariate dimension $D = 326$. The instances are classified into ‘musk’ (207 in total) or ‘non-musk’ (269 in total). Results are averaged over 100 random partitions of observations into $n = 250$ training and 226 test points.

Table 16 reports the average counts of selected variables and the average misclassification rates for all four datasets. In general, the SKDA methods tend to select too few variables. For dataset (i), LASSO and KNIFE include significantly more artificial variables than PLCR and PLLR. The LASSO performs exceptionally well in classification, suggesting linearity of $m(\boldsymbol{x})$. The misclassification rate of PLLR is close to that of LASSO, while the other methods are significantly worse. For datasets (ii), (iii) and (iv), PLLR has the lowest misclassification rates among all the methods. It is quite effective in screening out artificial variables, despite the relatively large number of variables it has selected for modelling $m(\boldsymbol{x})$. For all four datasets, NPLLR has notably higher misclassification rates than the other methods, thus confirming the importance of variable selection for classification problems.

Table 16

Example L3 (logistic regression) — number of selected artificial variables ($|\hat{\mathcal{A}}_a|$), number of selected variables ($|\hat{\mathcal{A}}|$) and misclassification rate (MR), averaged over 100 random partitions of dataset.

	(i) Wisconsin data			(ii) sonar data			(iii) ionosphere data			(iv) musk data		
Method	$\lvert\hat{\mathcal{A}}_a\rvert$	$\lvert\hat{\mathcal{A}}\rvert$	MR	$\lvert\hat{\mathcal{A}}_a\rvert$	$\lvert\hat{\mathcal{A}}\rvert$	MR	$\lvert\hat{\mathcal{A}}_a\rvert$	$\lvert\hat{\mathcal{A}}\rvert$	MR	$\lvert\hat{\mathcal{A}}_a\rvert$	$\lvert\hat{\mathcal{A}}\rvert$	MR
PLCR	0	9.66	0.122	0.13	6.39	0.273	0.02	5.99	0.19	1.21	6.23	0.120
PLLR	0.12	3.71	0.0482	1.84	18.0	**0.247**	0.21	9.56	**0.103**	0.81	8.80	**0.093**
KNIFE	2.02	10.5	0.0734	2.82	7.36	0.308	0	2.35	0.199	4.33	3.88	0.182
SKDA-bayes	0	1.81	0.0766	0	2.03	0.291	0	2	0.128	0	4.06	0.110
SKDA-mle	0	1.97	0.0700	0	1.98	0.277	0	2	0.111	0	7.23	0.105
LASSO	1.88	9.5	**0.0345**	0.95	9.48	0.261	1.14	9.99	0.169	8.23	36	0.112
NPLLR	0	30	0.0939	0	60	0.417	0	35	0.478	0	226	0.435

8.4. Quantile estimation of fitted regression function

Inference about $m(\boldsymbol{x})$ often entails estimation of quantiles of the sampling distribution of $\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})$. The asymptotic normality result derived in Theorem 2(iii) suggests a plug-in estimator of the γth-quantile $Q_\gamma(\boldsymbol{x})$ of $\hat{m}(\boldsymbol{x})-m(\boldsymbol{x})$, in which the unknown quantities are replaced by consistent estimators. Specifically, we estimate $\mathcal{A}$ by $\hat{\mathcal{A}}$ and $\nabla_{d,d}m$ by local cubic polynomial fitting with a Gaussian kernel and a bandwidth fixed by cross validation. In the case of mean regression, $\sigma(\boldsymbol{x})^2$ is estimated by a normalised weighted residual sum of squares (Fan & Gijbels, 1996), and $f_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$ by a kernel density estimator $\hat{f}_{\hat{\mathcal{A}}}(\boldsymbol{x}_{\hat{\mathcal{A}}})$ based on a Gaussian kernel and a bandwidth fixed by Silverman's rule. In the case of quantile regression, the conditional density of Y at $m(\boldsymbol{x})$, given $\boldsymbol{X} = \boldsymbol{x}$, is estimated by a similar kernel density estimator of the joint density of $(\boldsymbol{X}_{\hat{\mathcal{A}}}, Y)$ at $\big(\boldsymbol{x}_{\hat{\mathcal{A}}}, \hat{m}(\boldsymbol{x})\big)$, divided by the marginal density estimator $\hat{f}_{\hat{\mathcal{A}}}(\boldsymbol{x}_{\hat{\mathcal{A}}})$. Denote by $\hat{Q}_\gamma(\boldsymbol{x})$ a generic plug-in estimator of $Q_\gamma(\boldsymbol{x})$ constructed using the above procedure. We consider three choices of $\hat{Q}_\gamma(\boldsymbol{x})$, with $\hat{\mathcal{A}}$ determined respectively by PLLR, KNIFE and LASSO. As a benchmark we include also an oracle plug-in estimator $\hat{O}_\gamma(\boldsymbol{x})$, constructed using the correct active set $\mathcal{A}$. For the PLLR estimate, $\hat{m}(\boldsymbol{x})$ is calculated using the bandwidths output by the PLLR algorithm. The other three estimators calculate $\hat{m}(\boldsymbol{x})$ by local linear regression on $\boldsymbol{X}_{\hat{\mathcal{A}}}$, or $\boldsymbol{X}_{\mathcal{A}}$ in the oracle case, with bandwidths determined by cross validation. In each simulation, 100 randomly generated test points $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{100}$ are sorted such that $m(\boldsymbol{Z}_{(1)}) \le \cdots \le m(\boldsymbol{Z}_{(100)})$. For better insight into the peformance of the quantile estimators, the mean squared error of each $\hat{Q}_\gamma$, $\mathrm{MSE}(\hat{Q}_\gamma)$, is estimated on four different regions by averaging over 100 independent replicates of $25^{-1}\sum_{i=25(j-1)+1}^{25j}\big\{\hat{Q}_\gamma(\boldsymbol{Z}_{(i)}) - Q_\gamma(\boldsymbol{Z}_{(i)})\big\}^2$, $j = 1, 2, 3, 4$. Thus, the jth region contains the 25 test points associated with the jth smallest quarter of the $m(\boldsymbol{Z}_i)$ values. The true quantile $Q_\gamma(\cdot)$ is approximated by Monte Carlo simulation.

Figure 2 displays the mean squared errors $\mathrm{MSE}(\hat{O}_\gamma)$ and $\mathrm{MSE}(\hat{Q}_\gamma)$, for $\gamma =$ 0.05, 0.25, 0.50, 0.75, 0.95 and for the $\rho = 0$ cases of Examples M1 (mean regression), Q4 (median regression) and L1 (logistic regression). For the case of logistic regression, extreme values of the estimated probabilities $\hat{p}(\boldsymbol{Z}_i)$ may lead to exponentially large variance estimates, in which case the estimated variance

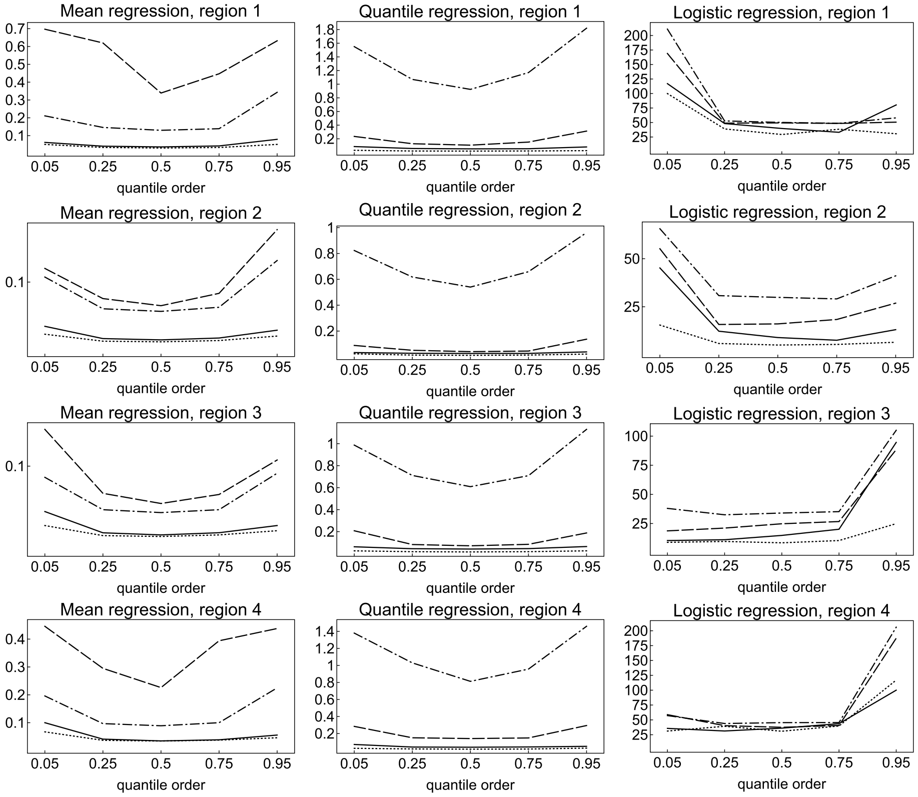

FIG 2. *Mean squared errors of oracle estimator $\hat{O}_\gamma$ (dotted) and estimators $\hat{Q}_\gamma$ of γth-quantile Q_γ, for $\gamma = 0.05, 0.25, 0.50, 0.75, 0.95$, with $\hat{Q}_\gamma$ constructed using variables selected by PLLR (solid), KNIFE (dashed) and LASSO (dot-dashed). Results for region j are averaged over test points corresponding to jth smallest quarter of regression function values.*

is truncated within the interval $[-100, 100]$. It is unsurprising that the more accurate is the estimated active set $\hat{\mathcal{A}}$, the better is the normal approximation to the distribution of $\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})$, hence the more accurate is $\hat{Q}_\gamma(\boldsymbol{x})$. This explains the superiority of the PLLR quantile estimator over the other estimators and its close resemblance to the oracle. The mean squared errors over the two extreme regions 1 and 4 are much larger than those over the middle two regions. In most cases the error is smallest at $\gamma = 0.5$ and increases gradually as γ deviates from 0.5 on either side. It may be of interest to note that the errors for logistic regression in regions 1 and 4 are skewed. Inspection of the distributions of $\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x})$ indicates that heavy tails appear on the left for $\boldsymbol{x}$ in region 1 and on the right for $\boldsymbol{x}$ in region 4, rendering normal approximation relatively poor for $\gamma = 0.05$ and 0.95, respectively.

9. Conclusion

We have proposed a new simultaneous variable selection and estimation method for nonparametric regression under a general convex loss function, and presented a coordinate descent algorithm for its implementation. The method is proved, under ultra-high dimensions, to yield selection consistency and estimate the true regression function at an oracle convergence rate. Our numerical examples evince that the new method is superior to existing methods in most cases except under additive models which may favour methods that explicitly exploit the additive structure. We have also derived the asymptotic distribution of the regression estimator and investigated empirically the accuracy of quantile estimators constructed by asymptotic approximation.

We may consider replacing the linear polynomial in (3.1) by a polynomial of a higher degree when formulating the objective function in (3.1). Extension of our proofs to this setting suggests a faster convergence rate for the estimator $\hat{m}(\boldsymbol{x})$, which, however, entails prohibitive computational expenses even when D is only moderately large.

Appendix A: Proof of Theorems 1 and 2

A.1. Preliminary lemmas

In what follows we denote by K_0 a sufficiently large positive constant, by k_0 a sufficiently small positive constant, and by $\{\epsilon_n\}$ a sequence of positive constants converging to zero at a sufficiently slow rate. The actual values of K_0, k_0 and $\{\epsilon_n\}$ may vary from occasion to occasion. For any sequences of random variables $\{W_n\}$ and $\{W_n'\}$, write $W_n' = \Omega_p(W_n)$ if $W_n = O_p(W_n')$ and $W_n' = O_p(W_n)$. The notation $\Omega(\cdot)$ is defined analogously for sequences of non-random constants.

Let $\mathcal{B} \subset \{1,\ldots,D\}$ be an arbitrary index set, possibly depending on n, such that $|\mathcal{B}| \le b_n \triangleq a_0 \log n/\log\log n$, for some constant $a_0 \in (0, 2/5)$. Let $a_n = k_0 n^{4/(4+b_n)}$ and $\bar{h} \triangleq \|\boldsymbol{h}_{\mathcal{B}}\|_\infty$. Consider a bandwidth vector $\boldsymbol{h} = (h_1,\ldots,h_D)^\top \in \mathscr{H}_{n,\mathcal{B}}$, where

$$\mathscr{H}_{n,\mathcal{B}} = \Big\{\boldsymbol{h} \in (0,\infty]^D : \bar{h} \le \epsilon_n b_n^{-3/2} \vee \epsilon_n/\omega^2,\ n\prod_{d\in\mathcal{B}} h_d \ge a_n, \\ \big(\min_{d\in\mathcal{B}^c} h_d\big)^{-1} \le K_0\Big\}.$$

Define, for $\mathcal{S} \subsetneq \{1,\ldots,n\}$, $d,d' \in \{0\}\cup\mathcal{B}$ and $\boldsymbol{x} \in \mathbb{R}^D$,

$$\kappa_{d,d'}^{-\mathcal{S}}(\boldsymbol{x}) = (nh_d h_{d'})^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i\notin\mathcal{S}} v(\boldsymbol{X}_i)V_i^{(d)}(\boldsymbol{x})V_i^{(d')}(\boldsymbol{x})K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x}),$$

where $h_0 \triangleq 1$ and $\big(V_i^{(0)}(\boldsymbol{x}), V_i^{(1)}(\boldsymbol{x}),\ldots,V_i^{(D)}(\boldsymbol{x})\big) \triangleq \big(1,(\boldsymbol{X}_i-\boldsymbol{x})^\top\big)$. Write $\kappa_{d,d'} = \kappa_{d,d'}^{-\emptyset}$ for brevity. Denote by $\big[\breve{\kappa}_{d,d'}(\boldsymbol{x})\big]_{d,d'\in\{0\}\cup\mathcal{B}}$ the inverse of the matrix $\big[\kappa_{d,d'}(\boldsymbol{x})\big]_{d,d'\in\{0\}\cup\mathcal{B}}$ so that $\sum_{j\in\{0\}\cup\mathcal{B}} \breve{\kappa}_{d,j}(\boldsymbol{x})\kappa_{d',j}(\boldsymbol{x}) = \mathbf{1}\{d=d'\}$, for any

$d, d' \in \{0\} \cup \mathcal{B}$. Define, for any function g on $\mathbb{R}^D$,

$$\mathcal{K}^r_{\mathcal{B}}(\boldsymbol{x}; g) = \int g(\boldsymbol{x}_{\mathcal{B}}, \boldsymbol{x}_{\mathcal{B}^c} + \boldsymbol{u}_{\mathcal{B}^c}) \prod_{d \in \mathcal{B}^c} K(u_d/h_d)^r d\boldsymbol{u}_{\mathcal{B}^c}.$$

Then we have

$$\mathbb{E}\kappa_{d,d'}(\boldsymbol{x}) = \begin{cases} \mathcal{K}^1_{\mathcal{B}}(\boldsymbol{x}; fv)\mu_{1,2\,\mathbf{1}\{d>0\}}\big\{1 + O(\bar{h})\big\}, & d = d', \\ O(\bar{h}), & d \neq d'. \end{cases}$$

Using Bernstein's inequality and boundedness of $v(\cdot)$, we have, for any $t \in (0, k_0)$ and any $d, d' \in \{0\} \cup \mathcal{B}$,

$$\mathbb{P}\big(|\kappa_{d,d'}(\boldsymbol{x}) - \mathbb{E}\kappa_{d,d'}(\boldsymbol{x})| \geq K_0\sqrt{t}\big) \leq 2e^{-nt\prod_{j\in\mathcal{B}} h_j}. \tag{S.1}$$

Define the event

$$\tilde{\mathcal{E}}_1 = \bigcap_{\mathcal{B}: |\mathcal{B}| \leq b_n} \bigcap_{d,d' \in \{0\} \cup \mathcal{B}} \Big\{ \sup_{\boldsymbol{h} \in \mathscr{H}_{n,\mathcal{B}}} |\kappa_{d,d'}(\boldsymbol{x}) - \mathbb{E}\kappa_{d,d'}(\boldsymbol{x})| \leq \epsilon_n/(|\mathcal{B}| \log |\mathcal{B}|) \Big\}.$$

It then follows by (S.1) that

$$\mathbb{P}(\tilde{\mathcal{E}}_1^c) \leq 2b_n^2 D^{b_n} e^{-k_0 a_n \epsilon_n^2/(b_n \log b_n)^2}. \tag{S.2}$$

Note that $\tilde{\mathcal{E}}_1$ implies that $\breve{\kappa}_{d,d'}(\boldsymbol{x}) \leq K_0$ if $d = d'$ and $\breve{\kappa}_{d,d'}(\boldsymbol{x}) \leq K_0\epsilon_n/(|\mathcal{B}| \log |\mathcal{B}|)$ if $d \neq d'$.

Write for brevity $\hat{\beta}_d(\boldsymbol{x}; \boldsymbol{h}) = \hat{\beta}_d^{-\emptyset}(\boldsymbol{x}; \boldsymbol{h})$, which solves the system of equations

$$\begin{aligned}(nh_d)^{-1}\Big(\prod_{j \in \mathcal{B}^c} h_j\Big) &\sum_{i=1}^n V_i^{(d)}(\boldsymbol{x}) K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) q_1\big(Y_i, \beta_0 + (\boldsymbol{X}_i - \boldsymbol{x})^\top \boldsymbol{\beta}\big) \\ &= -\,C_n(h_d)|\beta_d|^{\alpha-1} e(\beta_d)(nh_d)^{-1}\Big(\prod_{j \in \mathcal{B}^c} h_j\Big) \sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}), \\ &\qquad d = 0, \ldots, D, \end{aligned} \tag{S.3}$$

for $(\beta_0, \boldsymbol{\beta}^\top)$, where $e(a) = \mathbf{1}\{a > 0\} - \mathbf{1}\{a < 0\}$ for $a \neq 0$ and $|e(0)| \leq 1$. For ease of proof, we henceforth consider, without loss of generality, a bounded solution space for $(\beta_0, \boldsymbol{\beta}^\top)$ with $|\beta_0| \vee \|\boldsymbol{\beta}\|_\infty \leq K_0$.

Define the events $\mathcal{E}_0 = \{\max_{1 \leq i \leq n} L_0(Y_i) \leq K_0 \log n\}$,

$$\tilde{\mathcal{E}}_2 = \bigcap_{\mathcal{B}: |\mathcal{B}| < b_n} \bigcap_{d \in \mathcal{B}^c} \Big\{ \sup_{\boldsymbol{h} \in \mathscr{H}_{n,\mathcal{B}}} |\hat{\beta}_d(\boldsymbol{x}; \boldsymbol{h})| \leq K_0 n^{-1} D^{-1} \Big\},$$

$$\tilde{\mathcal{E}}_3 = \bigcap_{\mathcal{B}: |\mathcal{B}| < b_n} \Big\{ \sup_{\boldsymbol{h} \in \mathscr{H}_{n,\mathcal{B}}} |r_{\mathcal{B}}(\boldsymbol{x}; \boldsymbol{h}) - \hat{\beta}_0(\boldsymbol{x}; \boldsymbol{h})| \leq \epsilon_n \Big\},$$

where

$$r_{\mathcal{B}}(\boldsymbol{x};\boldsymbol{h}) = \operatorname*{argmin}_{\beta_0\in\mathbb{R}} \int f(\boldsymbol{x}_{\mathcal{B}},\boldsymbol{x}_{\mathcal{B}^c}+\boldsymbol{u}_{\mathcal{B}^c})\mathbb{E}\big[L(Y,\beta_0)\big|\boldsymbol{X}$$
$$= (\boldsymbol{x}_{\mathcal{B}},\boldsymbol{x}_{\mathcal{B}^c}+\boldsymbol{u}_{\mathcal{B}^c})\big] \prod_{d\in\mathcal{B}^c} K(u_d/h_d)\, d\boldsymbol{u}_{\mathcal{B}^c}.$$

Note that $r_{\mathcal{B}}(\boldsymbol{x};\boldsymbol{h}) = m(\boldsymbol{x})$ for $\mathcal{B}\supset\mathcal{A}$. If we set $\mathcal{B}=\mathcal{N}(\boldsymbol{c})$ and $\boldsymbol{h}_{\mathcal{N}(\boldsymbol{c})^c}=\boldsymbol{c}_{\mathcal{N}(\boldsymbol{c})^c}$, then $r_{\mathcal{B}}(\boldsymbol{x};\boldsymbol{h})$ reduces to $r(\boldsymbol{x};\boldsymbol{c})$ as defined in (5.1). Note also that there exists by (A6) a sufficiently large $K_0>0$ such that $\mathbb{P}(\mathcal{E}_0)\to 1$.

Denote, for brevity, by $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, $\mathcal{V}(\boldsymbol{x})$ and $\mathcal{V}_i(\boldsymbol{x})$ the $(|\mathcal{B}|+1)$-dimensional subvectors $\big(\beta_d : d\in\{0\}\cup\mathcal{B}\big)$, $\big(\hat\beta_d(\boldsymbol{x};\boldsymbol{h}) : d\in\{0\}\cup\mathcal{B}\big)$, $\big(X^{(d)}-x_d : d\in\{0\}\cup\mathcal{B}\big)$ and $\big(V_i^{(d)}(\boldsymbol{x}) : d\in\{0\}\cup\mathcal{B}\big)$, respectively.

The proof of Theorem 1 is lengthy so we give below a short navigation. Lemma 1 constructs a uniform probability bound on each component of the estimator (3.1), which holds for all $|\mathcal{B}|<b_n$ and, in particular, for $\mathcal{B}=\mathcal{A}$ or $\hat{\mathcal{A}}$. The results enable us to study the cross validation error of $\hat\beta_0^{-\mathcal{S}}(\boldsymbol{x};\boldsymbol{h})$ for a general class of bandwidths $\boldsymbol{h}$, and show that $\hat{\mathcal{A}}\supset\mathcal{A}$ with large probability by contrasting the error incurred under $\mathcal{B}\supset\mathcal{A}$ with that incurred under $\mathcal{B}\not\supset\mathcal{A}$, the latter being significantly bigger. Lemma 2 scrutinises the remainder terms of the estimation error, $\hat\beta_0(\boldsymbol{x};\boldsymbol{h})-m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$, under the assumption $\mathcal{B}\supset\mathcal{A}$. A precise expansion for the cross validation error is then established in Lemma 3. The above results are finally combined to prove Theorem 1.

Lemma 1. *Assume the conditions of Theorem 1. Then we have*

$$\mathbb{P}(\mathcal{E}_0\cap\tilde{\mathcal{E}}_2^c) \le 2D^{2+b_n}e^{-k_0 n\prod_{d\in\mathcal{B}}h_d} \le 2D^{2+b_n}e^{-k_0 a_n}, \tag{S.4}$$
$$\mathbb{P}(\mathcal{E}_0\cap\tilde{\mathcal{E}}_2\cap\tilde{\mathcal{E}}_3^c) \le 2D^{b_n}e^{-k_0 a_n\epsilon_n^2/\log n}. \tag{S.5}$$

The first result (S.4) asserts that $\big\{|\hat\beta_d(\boldsymbol{x};\boldsymbol{h})| : d\in\mathcal{B}^c\big\}$ are uniformly bounded by a negligibly small sequence with large probability, so that unselected variables make only negligibly small contributions to the local linear expansion. The second result (S.5) states that $\hat\beta_0(\boldsymbol{x};\boldsymbol{h})$ has a leading term $r_{\mathcal{B}}(\boldsymbol{x};\boldsymbol{h})$ with large probability under an arbitrary choice of $(\mathcal{B},\boldsymbol{h})$.

Proof of Lemma 1. Let $M_\beta(\boldsymbol{x}) = \max\big\{|\hat\beta_d(\boldsymbol{x};\boldsymbol{h})| : d\in\mathcal{B}^c,\ |\mathcal{B}|<b_n\big\}$, so that $\tilde{\mathcal{E}}_2 = \big\{M_\beta(\boldsymbol{x})\le K_0 n^{-1}D^{-1}\big\}$. Note that

$$\mathbb{P}\Big(n^{-1}\prod_{j\in\mathcal{B}^c}h_j\Big|\sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})L\big(Y_i,0\big)-\mathbb{E}\sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})L\big(Y_i,0\big)\Big|\ge K_0\sqrt{t}\Big)$$
$$\le 2e^{-nt\prod_{j\in\mathcal{B}}h_j}.$$

The objective value of (3.1) at $(\beta_0,\boldsymbol{\beta}^\top) = \mathbf{0}^\top$ exceeds K_0 with probability bounded by $e^{-k_0 n\prod_{j\in\mathcal{B}}h_j}\le e^{-k_0 a_n}$. It follows by minimality of $\big\{\hat\beta_d(\boldsymbol{x};\boldsymbol{h}) :$

$d = 0,\ldots,D\big\}$ that the penalty $\sum_{d=1}^{D} C_n(h_d)|\hat{\beta}_d(\boldsymbol{x};\boldsymbol{h})|^{\alpha} \le \sum_{i=1}^{n} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x})L(Y_i,0)$. It then follows that $M_\beta(\boldsymbol{x})^{\alpha} \le K_0 n^{-\alpha\vee 2+1}D^{-\alpha\vee 2+1}$. In particular, (S.4) holds for $\alpha = 1$. For $\alpha > 1$, we have, on $\mathcal{E}_0$, that for any fixed $d \in \mathcal{B}^c$,

$$\Big|(nh_d)^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^{n} V_i^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})q_1\big(Y_i,\sum_{d'=0}^{D} V_i^{(d')}(\boldsymbol{x})\hat{\beta}_{d'}(\boldsymbol{x};\boldsymbol{h})\big)\Big|$$
$$\le K_0(\log n)\sum_{d'=0}^{D}|\hat{\beta}_{d'}(\boldsymbol{x};\boldsymbol{h})|T_{1,d,d'} + |T_{2,d}|,$$

where

$$T_{1,d,d'} = n^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^{n}\big|V_i^{(d)}(\boldsymbol{x})V_i^{(d')}(\boldsymbol{x})\big|K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x}),$$
$$T_{2,d} = n^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^{n} V_i^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})q_1(Y_i,0).$$

Using Bernstein's inequality, we have

$$\mathbb{P}\big(|T_{2,d}-\mathbb{E}T_{2,d}|\ge K_0\sqrt{t}\big)\vee\mathbb{P}\big(|T_{1,d,d'}-\mathbb{E}T_{1,d,d'}|\ge K_0\sqrt{t}\big)\le 2e^{-nt\prod_{j\in\mathcal{B}}h_j},$$

with $\mathbb{E}T_{1,d,d'} = O(1)$ and $\mathbb{E}T_{2,d} = O(1)$. It follows that $\max\big\{T_{1,d,d'}\vee|T_{2,d}| : 0\le d'\le D,\, d\in\mathcal{B}^c\big\}\le K_0$ with probability greater than $1-2D^2e^{-k_0 n\prod_{j\in\mathcal{B}}h_j}$. Assuming $\max\big\{T_{1,d,d'}\vee|T_{2,d}| : 0\le d'\le D,\, d\in\mathcal{B}^c\big\}\le K_0$, it follows from (S.3) that

$$M_\beta(\boldsymbol{x})^{\alpha-1}\le K_0\big\{n^{-(\alpha\vee 2-1)}D^{-(\alpha\vee 2-1)} + n^{-(\alpha\vee 2-1)}D^{-(\alpha\vee 2-2)}(\log n)M_\beta(\boldsymbol{x})\big\}. \tag{S.6}$$

For $\alpha\in(1,2]$, (S.6) reduces to $M_\beta(\boldsymbol{x})^{\alpha-1}\le K_0\big\{n^{-1}D^{-1}+n^{-1}(\log n)M_\beta(\boldsymbol{x})\big\}$, so that $M_\beta(\boldsymbol{x})\le K_0 n^{-1}D^{-1}$. For $\alpha > 2$, we have by (S.6) either $M_\beta(\boldsymbol{x})\le(2K_0)^{1/(\alpha-1)}n^{-1}D^{-1}$ or $M_\beta(\boldsymbol{x})\le(2K_0)^{1/(\alpha-2)}n^{-(\alpha-1)/(\alpha-2)}D^{-1}\log n$. It follows that (S.4) also holds for any $\alpha > 1$.

Let $T_3(\beta_0,\boldsymbol{\beta}) = n^{-1}\big(\prod_{j\in\mathcal{B}^c}h_j\big)\sum_{i=1}^{n}K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})L\big(Y_i,\sum_{d=0}^{D}V_i^{(d)}(\boldsymbol{x})\beta_d\big)$. Then we have, for $(\beta_0,\boldsymbol{\beta})$ satisfying $\|\boldsymbol{\beta}_{\mathcal{B}^c}\|_\infty\le K_0n^{-1}D^{-1}$,

$$\big|T_3(\beta_0,\boldsymbol{\beta})-T_{3,1}(\beta_0)-T_{3,2}(\boldsymbol{\theta})\big|\le K_0n^{-1}\prod_{j\in\mathcal{B}}h_j^{-1}, \tag{S.7}$$

where

$$T_{3,1}(\beta_0) = n^{-1}\Big(\prod_{j\in\mathcal{B}^c}h_j\Big)\sum_{i=1}^{n}K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})L\big(Y_i,\beta_0\big),$$
$$T_{3,2}(\boldsymbol{\theta}) = n^{-1}\Big(\prod_{j\in\mathcal{B}^c}h_j\Big)\sum_{i=1}^{n}K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})\Big\{L\big(Y_i,\mathcal{V}_i(\boldsymbol{x})^{\top}\boldsymbol{\theta}\big)-L\big(Y_i,\beta_0\big)\Big\}.$$

By the Talagrand's concentration inequality (Talagrand, 1996), we have

$$\mathbb{P}\Big(\sup_{|\beta_0|\le K_0} |T_{3,1}(\beta_0) - \mathbb{E}T_{3,1}(\beta_0)| \ge K_0\sqrt{t}\Big) \le K_0 e^{-nt\prod_{j\in\mathcal{B}} h_j/\log n}, \tag{S.8}$$

with

$$\begin{aligned}\mathbb{E}T_{3,1}(\beta_0) = M_1 \int f(\boldsymbol{x}_{\mathcal{B}}, (\boldsymbol{x}+\boldsymbol{u})_{\mathcal{B}^c})\mathbb{E}\big[L(Y,\beta_0)\big|\boldsymbol{X} = (\boldsymbol{x}_{\mathcal{B}}, (\boldsymbol{x}+\boldsymbol{u})_{\mathcal{B}^c})\big]& \\ \times \prod_{d\in\mathcal{B}^c} K(u_d/h_d)\, d\boldsymbol{u}_{\mathcal{B}^c} + O\Big(\sum_{d\in\mathcal{B}} h_d + n^{-1}\Big)&\end{aligned} \tag{S.9}$$

for some constant $M_1 > 0$ not depending on β_0. We have also

$$\begin{aligned}&\big|T_{3,2}(\boldsymbol{\theta})\big| \\ &\le n^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x})\big|q_1(Y_i,0)\big|\sum_{d\in\mathcal{B}}|V_i^{(d)}(\boldsymbol{x})\beta_d| \\ &\quad + K_0 n^{-1}\log n\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x})\sum_{d\in\{0\}\cup\mathcal{B}}\big|V_i^{(d)}(\boldsymbol{x})\beta_d\big|\sum_{d'\in\mathcal{B}}|V_i^{(d')}(\boldsymbol{x})\beta_{d'}| \\ &\le K_0 n^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^n K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x})\sum_{d\in\mathcal{B}}|V_i^{(d)}(\boldsymbol{x})|\big\{|q_1(Y_i,0)| + \|\mathcal{V}_i(\boldsymbol{x})\|_1\log n\big\},\end{aligned}$$

so that

$$\mathbb{P}\Big(\sup_{\|\boldsymbol{\theta}\|_\infty\le K_0} |T_{3,2}(\boldsymbol{\theta})| > K_0|\mathcal{B}|(\bar{h}\log n + \sqrt{t})\Big) \le 3|\mathcal{B}|^2 e^{-nt\prod_{j\in\mathcal{B}} h_j}. \tag{S.10}$$

It follows from (S.7)–(S.10) that $\mathbb{P}\big(|r_{\mathcal{B}}(\boldsymbol{x};\boldsymbol{h}) - \hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})| \ge \epsilon_n\big) \le K_0 e^{-k_0 a_n \epsilon_n^2/\log n}$, and hence (S.5) follows. $\blacksquare$

Next we examine in detail the remainder terms in the expansion for the estimator $\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})$, under the case $\mathcal{B} \supset \mathcal{A}$ where $\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})$ is consistent for $m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$.

Define, for $i = 1,\ldots,n$,

$$P_i(\boldsymbol{x},\boldsymbol{\theta}) = \int_0^1 (1-u)q_3\big(m(\boldsymbol{X}_i) - u\{m(\boldsymbol{X}_i) - \boldsymbol{\theta}^\top\mathcal{V}_i(\boldsymbol{x})\}\big)\,du$$

and $w_i = -q_1\big(Y_i, m(\boldsymbol{X}_i)\big)$. Define

$$\begin{aligned}q_n &= \sqrt{2\log(2|\mathcal{B}|)/n} + \Big(\prod_{j\in\mathcal{B}} h_j\Big)^{-1/2}\log(2|\mathcal{B}|)/n, \\ \Xi &= n^{-1} + n^{\alpha\vee 2}D^{(\alpha\vee 2)-1}\sum_{d\in\mathcal{B}} e^{-c/h_d}\end{aligned}$$

and

$$\boldsymbol{\beta}^* = \big(\beta_d^* : d\in\{0\}\cup\mathcal{B}\big) = \big(\nabla_d m(\boldsymbol{x}) : d\in\{0\}\cup\mathcal{B}\big).$$

Define, for $d \in \{0\} \cup \mathcal{B}$,

$$\nu_{d,i}(\boldsymbol{x}) = (nh_d)^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big) \sum_{d'\in\{0\}\cup\mathcal{B}} h_{d'}^{-1}\breve{\kappa}_{d,d'}(\boldsymbol{x}) V_i^{(d')}(\boldsymbol{x}) K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}),\ i = 1,\ldots,n,$$

$$R_d^*(\boldsymbol{x}) = h_d^{-1} \sum_{d_1,d_2\in\mathcal{B}} h_{d_1}h_{d_2} \sum_{d'\in\{0\}\cup\mathcal{B}} \breve{\kappa}_{d,d'}(\boldsymbol{x}) T_{4,d',d_1,d_2},$$

where, for $d_1, d_2 \in \mathcal{B}$ and $d' \in \{0\} \cup \mathcal{B}$,

$$\begin{aligned} T_{4,d',d_1,d_2} &= (nh_{d_1}h_{d_2}h_{d'})^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big) \\ &\quad \times \sum_{i=1}^n v(\boldsymbol{X}_i) V_i^{(d')}(\boldsymbol{x}) V_i^{(d_1)}(\boldsymbol{x}) V_i^{(d_2)}(\boldsymbol{x}) K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) R_{d_1,d_2,\boldsymbol{x}}(\boldsymbol{X}_i), \\ R_{d_1,d_2,\boldsymbol{x}}(\boldsymbol{y}) &= \int_0^1 (1-t)\nabla_{d_1,d_2} m_{\mathcal{A}}\left((1-t)\boldsymbol{x}_{\mathcal{A}} + t\boldsymbol{y}_{\mathcal{A}}\right) dt,\ \ \boldsymbol{y} \in \mathbb{R}^D. \end{aligned}$$

For $d \in \{0\} \cup \mathcal{B}$, define

$$\begin{aligned} Q_{1,d}(\boldsymbol{x},\boldsymbol{\theta}) &= \sum_{\ell=1}^n \nu_{d,\ell}(\boldsymbol{x}) \Big\{ -q_1\big(Y_\ell, \boldsymbol{\theta}^\top \mathcal{V}_\ell(\boldsymbol{x})\big) + q_1\big(Y_\ell, \boldsymbol{\beta}^{*\top}\mathcal{V}_\ell(\boldsymbol{x})\big) \\ &\qquad + \eta\big(\boldsymbol{\theta}^\top \mathcal{V}_\ell(\boldsymbol{x})\big|\boldsymbol{X}_\ell\big) - \eta\big(\boldsymbol{\beta}^{*\top}\mathcal{V}_\ell(\boldsymbol{x})\big|\boldsymbol{X}_\ell\big)\Big\}, \\ Q_{2,d}(\boldsymbol{x}) &= \sum_{\ell=1}^n \nu_{d,\ell}(\boldsymbol{x}) \Big\{ q_1\big(Y_\ell, m(\boldsymbol{X}_\ell)\big) - q_1\big(Y_\ell, \boldsymbol{\beta}^{*\top}\mathcal{V}_\ell(\boldsymbol{x})\big) - \eta\big(m(\boldsymbol{X}_\ell)\big|\boldsymbol{X}_\ell\big) \\ &\qquad + \eta\big(\boldsymbol{\beta}^{*\top}\mathcal{V}_\ell(\boldsymbol{x})\big|\boldsymbol{X}_\ell\big)\Big\}, \\ Q_{3,d}(\boldsymbol{x},\boldsymbol{\theta}) &= -\sum_{\ell=1}^n \nu_{d,\ell}(\boldsymbol{x}) P_\ell(\boldsymbol{x},\boldsymbol{\theta}) \big\{ m(\boldsymbol{X}_\ell) - \boldsymbol{\theta}^\top \mathcal{V}_\ell(\boldsymbol{x})\big\}^2. \end{aligned}$$

Define also $Q(\boldsymbol{x}) = Q_{1,0}(\boldsymbol{x},\hat{\boldsymbol{\theta}}) + Q_{2,0}(\boldsymbol{x}) + Q_{3,0}(\boldsymbol{x},\hat{\boldsymbol{\theta}})$.

Lemma 2. *Assume the conditions of Theorem 1, $\mathcal{A} \subset \mathcal{B}$ and that $\mathcal{E}_0 \cap \tilde{\mathcal{E}}_1 \cap \tilde{\mathcal{E}}_2 \cap \tilde{\mathcal{E}}_3$ holds. Then $\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})$ admits an expansion satisfying*

$$\Big|\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h}) - m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}) - \sum_{i=1}^n \nu_{0,i}(\boldsymbol{x}) w_i - R_0^*(\boldsymbol{x}) - Q(\boldsymbol{x})\Big| \le K_0 \Xi.$$

Moreover, up to an additive error bounded by $K_0\big\{\mathbb{P}(\mathcal{E}_0^c) + \sum_{j=1}^3 \mathbb{P}(\tilde{\mathcal{E}}_j^c)\big\} = o(1)$, we have $\mathbb{E}R_0^*(\boldsymbol{x}) = O\big(\omega\bar{h}^2\big)$, which can be sharpened to $\Omega\big(\omega\bar{h}^2\big)$ under the stronger condition (A2'), and, for $t > 0$,

$$\mathbb{P}\big(|R_0^*(\boldsymbol{x}) - \mathbb{E}R_0^*(\boldsymbol{x})| \ge \omega\bar{h}^2 K_0\sqrt{t}\big) \le 3|\mathcal{B}| e^{-nt\prod_{j\in\mathcal{B}} h_j},$$

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n \nu_{0,i}(\boldsymbol{x}) w_i\Big| \ge K_0\sqrt{t}\Big) \le 3e^{-nt\prod_{j\in\mathcal{B}} h_j},$$

$$\begin{aligned}
&\mathbb{P}\Big(|Q_{1,0}(\boldsymbol{x},\hat{\boldsymbol{\theta}})| \geq K_0|\mathcal{B}|\big(\omega\bar{h}^2+\sqrt{t}\big)\big\{K_0 q_n(\log n)\prod_{j\in\mathcal{B}} h_j^{-1/2}+\sqrt{t}\big\}\Big)\\
&\qquad \leq K_0|\mathcal{B}|e^{-nt\prod_{j\in\mathcal{B}} h_j},\\
&\mathbb{P}\big(|Q_{2,0}(\boldsymbol{x})| \geq \omega\bar{h}^2 K_0\sqrt{t}\big) \leq 3e^{-nt\prod_{j\in\mathcal{B}} h_j},\\
&\mathbb{P}\big(|Q_{3,0}(\boldsymbol{x},\hat{\boldsymbol{\theta}})| \geq |\mathcal{B}|^2(K_0\omega\bar{h}^2+\sqrt{t})^2\big) \leq K_0|\mathcal{B}|e^{-nt\prod_{j\in\mathcal{B}} h_j}.
\end{aligned}$$

Lemma 2 expresses the estimation error $\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})-m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})$ as a sum of three terms, up to an additive error of a small order. The first two terms, $\sum_{i=1}^n \nu_{0,i}(\boldsymbol{x})w_i$ and $R_0^*(\boldsymbol{x})$, account for the variance and bias of local linear regression, respectively. The third term $Q(\boldsymbol{x})$ has a smaller order than the above two. The lemma also establishes probability bounds for these three terms, which are useful for proving the next lemma and Theorem 2.

Proof of Lemma 2. Note that, for $i=1,\ldots,n$,

$$\begin{aligned}
\eta\big(\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})\big|\boldsymbol{X}_i\big) = \eta\big(m(\boldsymbol{X}_i)\big|\boldsymbol{X}_i\big) + v(\boldsymbol{X}_i)\big\{\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})-m(\boldsymbol{X}_i)\big\}&\\
+P_i(\boldsymbol{x},\hat{\boldsymbol{\theta}})\big\{\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})-m(\boldsymbol{X}_i)\big\}^2.&\qquad\text{(S.11)}
\end{aligned}$$

It follows from (S.3), (S.4) and (S.11) that $\sum_{d'\in\{0\}\cup\mathcal{B}} h_{d'}\kappa_{d,d'}(\boldsymbol{x})\hat{\beta}_{d'}(\boldsymbol{x};\boldsymbol{h})$ admits an expansion

$$\begin{aligned}
&(nh_d)^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\sum_{i=1}^n V_i^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})\\
&\times\Big[v(\boldsymbol{X}_i)m(\boldsymbol{X}_i)-q_1(Y_i,m(\boldsymbol{X}_i))-\big\{q_1\big(Y_i,\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})\big)-q_1\big(Y_i,m(\boldsymbol{X}_i)\big)\big\}\\
&\quad+\big\{\eta\big(\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})\big|\boldsymbol{X}_i\big)-\eta\big(m(\boldsymbol{X}_i)|\boldsymbol{X}_i\big)\big\}-P_i(\boldsymbol{x},\hat{\boldsymbol{\theta}})\big\{m(\boldsymbol{X}_i)-\hat{\boldsymbol{\theta}}^\top\mathcal{V}_i(\boldsymbol{x})\big\}^2\Big]\\
&\hspace{30em}\text{(S.12)}
\end{aligned}$$

up to an error bounded by $K_0h_d^{-1}\Xi$. By substituting

$$\begin{aligned}
m(\boldsymbol{X})-\hat{\boldsymbol{\theta}}^\top\mathcal{V}(\boldsymbol{x}) = &\sum_{d_1,d_2\in\mathcal{B}}(X^{(d_1)}-x_{d_1})(X^{(d_2)}-x_{d_2})R_{d_1,d_2,\boldsymbol{x}}(\boldsymbol{X})\\
&+\sum_{d\in\{0\}\cup\mathcal{B}}\big\{\nabla_d m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})-\hat{\beta}_d(\boldsymbol{x};\boldsymbol{h})\big\}(X^{(d)}-x_d)
\end{aligned}$$

and inverting (S.12), we have, for $d\in\{0\}\cup\mathcal{B}$, that

$$\begin{aligned}
&\Big|\hat{\beta}_d(\boldsymbol{x};\boldsymbol{h})-\nabla_d m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})-\sum_{i=1}^n\nu_{d,i}(\boldsymbol{x})w_i-R_d^*(\boldsymbol{x})\\
&\qquad-Q_{1,d}(\boldsymbol{x},\hat{\boldsymbol{\theta}})-Q_2(\boldsymbol{x})-Q_{3,d}(\boldsymbol{x},\hat{\boldsymbol{\theta}})\Big| \leq K_0h_d^{-1}\Xi.\qquad\text{(S.13)}
\end{aligned}$$

The expansion for $\hat{\beta}_0(\boldsymbol{x};\boldsymbol{h})$ then follows by setting $d=0$ in (S.13).

In what follows we omit from all probability bounds an additive error $K_0\big\{\mathbb{P}(\mathcal{E}_0^c)+\sum_{j=1}^3\mathbb{P}(\tilde{\mathcal{E}}_j^c)\big\}=o(1)$ for simplicity. It can be deduced, for $d\in\{0\}\cup\mathcal{B}$, that

$$\mathbb{P}\Big((nh_d)^{-1}\Big(\prod_{j\in\mathcal{B}^c} h_j\Big)\Big|\sum_{i=1}^n V_i^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})w_i\Big|\geq K_0\sqrt{t}\Big)\leq 2e^{-nt\prod_{j\in\mathcal{B}}h_j}.$$

so that

$$\mathbb{P}\Big(h_d\Big|\sum_{i=1}^n \nu_{d,i}(\boldsymbol{x})w_i\Big|\geq K_0\sqrt{t}\Big)\leq 3e^{-nt\prod_{j\in\mathcal{B}}h_j}. \tag{S.14}$$

Similarly, we have, for $d\in\{0\}\cup\mathcal{B}$, that

$$\mathbb{P}\big(|Q_{2,d}(\boldsymbol{x})|\geq K_0\omega\bar{h}^2\sqrt{t}\big)\leq 3e^{-nt\prod_{j\in\mathcal{B}}h_j}. \tag{S.15}$$

Define, for $\boldsymbol{\theta}\in\mathbb{R}^{|\mathcal{B}|+1}$, $\delta(\boldsymbol{\theta})=\sum_{d'\in\{0\}\cup\mathcal{B}}h_{d'}|\beta_{d'}-\beta^*_{d'}|$. Following the notations of van de Geer (2008), define $\mathcal{F}_M\triangleq\{\boldsymbol{\theta}:\delta(\boldsymbol{\theta})\leq M\}$ and

$$\begin{aligned}\iota_{\boldsymbol{\theta},d}(\boldsymbol{X},Y)=h_d^{-1}\Big(\prod_{j\in\mathcal{B}^c}h_j\Big)\Big(\prod_{d'\in\{0\}\cup\mathcal{B}}h_{d'}^{1/2}\Big)&\\ \times(X^{(d)}-x_d)K_{\boldsymbol{h}}(\boldsymbol{X}-\boldsymbol{x})q_1\big(Y,\boldsymbol{\theta}^\top\mathcal{V}(\boldsymbol{x})\big).&\end{aligned}$$

We provide next a probability bound for $\sup_{\boldsymbol{\theta}\in\mathcal{F}_M}|Q_{1,d}(\boldsymbol{x},\boldsymbol{\theta})|$. A symmetrisation theorem (van der Vaart & Wellner, 1996) can be used to bound the mean of

$$\begin{aligned}&\boldsymbol{Z}_d(M)\\ &\triangleq\sup_{\boldsymbol{\theta}\in\mathcal{F}_M}\Big|n^{-1}\sum_{i=1}^n\big\{\iota_{\boldsymbol{\theta},d}(\boldsymbol{X}_i,Y_i)-\iota_{\boldsymbol{\beta}^*,d}(\boldsymbol{X}_i,Y_i)\big\}-\mathbb{E}\big[\iota_{\boldsymbol{\theta},d}(\boldsymbol{X},Y)-\iota_{\boldsymbol{\beta}^*,d}(\boldsymbol{X},Y)\big]\Big|\end{aligned}$$

by $2\,\mathbb{E}\sup_{\boldsymbol{\theta}\in\mathcal{F}_M}\big|n^{-1}\sum_{i=1}^n\{\iota_{\boldsymbol{\theta},d}(\boldsymbol{X}_i,Y_i)-\iota_{\boldsymbol{\beta}^*,d}(\boldsymbol{X}_i,Y_i)\}E_i\big|$, for a Rademacher sequence $\{E_i\}$ independent of $\mathscr{D}_n$, for all $d\in\{0\}\cup\mathcal{B}$. Applying the contraction theorem (Ledoux & Talagrand, 1991), we have

$$\begin{aligned}\mathbb{E}\boldsymbol{Z}_d(M)&\leq K_0n^{-1}(\log n)h_d^{-1}\Big(\prod_{j\in\mathcal{B}^c}h_j\Big)\Big(\prod_{j\in\mathcal{B}}h_j^{1/2}\Big)\\ &\qquad\times\mathbb{E}\sup_{\boldsymbol{\theta}\in\mathcal{F}_M}\sum_{i=1}^nE_iK_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})\sum_{d'\in\{0\}\cup\mathcal{B}}\big|V_i^{(d)}(\boldsymbol{x})V_i^{(d')}(\boldsymbol{x})(\beta_{d'}-\beta^*_{d'})\big|\\ &\leq K_0Mn^{-1}(\log n)h_d^{-1}\Big(\prod_{j\in\mathcal{B}^c}h_j\Big)\Big(\prod_{j\in\mathcal{B}}h_j^{1/2}\Big)\\ &\qquad\times\mathbb{E}\Big[\max_{d'\in\{0\}\cup\mathcal{B}}h_{d'}^{-1}\sum_{i=1}^nK_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})\big|V_i^{(d)}(\boldsymbol{x})V_i^{(d')}(\boldsymbol{x})\big|\Big]\\ &\leq K_0Mq_n\log n.\end{aligned}\tag{S.16}$$

The last inequality above follows from Lemma A.1 of van de Geer (2008). Applying the Bousquet inequality (Bousquet, 2002), we have

$$\mathbb{P}\big(\boldsymbol{Z}_d(M)\geq K_0Mq_n(K_0\log n+z\sqrt{1+\eta_nq_n\log n}+2z^2q_n\eta_n/3)\big)\leq e^{-nq_n^2z^2},$$

where $\eta_n = K_0 \prod_{j \in \mathcal{B}} h_j^{-1/2} \geq \sup_{1 \leq i \leq n, d \in \mathcal{B}, \boldsymbol{\theta} \in \mathcal{F}_{K_0}} \big| \iota_{\boldsymbol{\theta}, d}(\boldsymbol{X}_i, Y_i) \big|$. It follows that

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta} \in \mathcal{F}_M} |Q_{1,d}(\boldsymbol{x}, \boldsymbol{\theta})| \geq K_0 M \big\{ K_0 q_n (\log n) \prod_{j \in \mathcal{B}} h_j^{-1/2} + \sqrt{t} \big\} \Big) \leq 3|\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j}. \tag{S.17}$$

For quantile regression and $|\mathcal{A}| = O(1)$, it follows from Central Limit Theorem that

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta} \in \mathcal{F}_M} |Q_{1,d}(\boldsymbol{x}, \boldsymbol{\theta})| \geq K_0 M \sqrt{t} \Big) \leq 2e^{-nt \prod_{j \in \mathcal{B}} h_j}. \tag{S.18}$$

Similar arguments show that

$$\mathbb{P}(\sup_{\boldsymbol{\theta} \in \mathcal{F}_M} |Q_{3,d}(\boldsymbol{x}, \boldsymbol{\theta}) - \mathbb{E} Q_{3,d}(\boldsymbol{x}, \boldsymbol{\theta})| \geq K_0 \big\{ M^2 + \omega^2 \bar{h}^4 \big\} \sqrt{t} \big) \leq 3e^{-nt \prod_{j \in \mathcal{B}} h_j}, \tag{S.19}$$

with $\mathbb{E} Q_{3,d}(\boldsymbol{x}, \boldsymbol{\theta}) = O\big(M^2 + \omega^2 \bar{h}^4\big)$ for any $\boldsymbol{\theta} \in \mathcal{F}_M$. Note that, for any $d' \in \{0\} \cup \mathcal{B}$ and $d_1, d_2 \in \mathcal{B}$, $\mathbb{P}\big(|T_{4,d',d_1,d_2} - \mathbb{E} T_{4,d',d_1,d_2}| \geq K_0 \sqrt{t}\big) \leq 2e^{-nt \prod_{j \in \mathcal{B}} h_j}$, with $\mathbb{E} T_{4,d',d_1,d_2} = (1/2) \nabla_{d_1,d_2} m_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}}) \mathcal{K}_{\mathcal{B}}^1(\boldsymbol{x}; fv) \mu_{1,2} \, \mathbf{1}\{d_1 = d_2, \, d' = 0\} + O(\bar{h})$. It follows that, for $d \in \{0\} \cup \mathcal{B}$,

$$\mathbb{P}\big(|R_d^*(\boldsymbol{x}) - \mathbb{E} R_d^*(\boldsymbol{x})| \geq \omega \bar{h}^2 K_0 \sqrt{t}\big) \leq 3|\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j}. \tag{S.20}$$

Combining (S.13), (S.14), (S.15), (S.17), (S.19) and (S.20), we have

$$\mathbb{P}\Big(|\delta(\hat{\boldsymbol{\theta}})| \geq K_0 |\mathcal{B}| \big(\omega \bar{h}^2 + \sqrt{t}\big) + K_0 \Xi\Big) \leq K_0 |\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j}. \tag{S.21}$$

The probability bounds in the lemma then follow by substituting (S.21) into (S.17) and (S.19), and by setting $d = 0$ in (S.20), (S.14), (S.17), (S.15) and (S.19), respectively. ▮

Define, for $d \in \{0\} \cup \mathcal{B}$, $R_d^{*-\mathcal{S}_k}$, $\nu_{d,i}^{-\mathcal{S}_k}$ $(i \notin \mathcal{S}_k)$ and $Q^{-\mathcal{S}_k}$ to be the counterparts of R_d^*, $\nu_{d,i}$ and Q, respectively, evaluated on the delete-$\mathcal{S}_k$ sample. Define, for $i \in \mathcal{S}_k$, $a_i^{-\mathcal{S}_k}(\boldsymbol{h}) = \hat{\beta}_0^{-\mathcal{S}_k}(\boldsymbol{X}_i; \boldsymbol{h}) - m(\boldsymbol{X}_i)$ and $\alpha_i^{-\mathcal{S}_k}(\boldsymbol{h}) = L\big(Y_i, m(\boldsymbol{X}_i) + a_i^{-\mathcal{S}_k}(\boldsymbol{h})\big) - L\big(Y_i, m(\boldsymbol{X}_i)\big) - q_1\big(Y_i, m(\boldsymbol{X}_i)\big) a_i^{-\mathcal{S}_k}(\boldsymbol{h})$. Define

$$G_1(\boldsymbol{h}) = (2n_0)^{-1} \sum_{i \in \mathcal{S}_1} v(\boldsymbol{X}_i) \Big\{ R_0^{*-\mathcal{S}_1}(\boldsymbol{X}_i) + (1 - n_0/n)^{-1} \sum_{j \notin \mathcal{S}_1} \nu_{0,j}^{-\mathcal{S}_1}(\boldsymbol{X}_i) w_j$$
$$+ Q^{-\mathcal{S}_1}(\boldsymbol{X}_i) \Big\}^2,$$

$$G_2(\boldsymbol{h}) = n_0^{-1} \sum_{i \in \mathcal{S}_1} w_i a_i^{-\mathcal{S}_1}(\boldsymbol{h}) + n_0^{-1} \sum_{i \in \mathcal{S}_1} \Big\{ \alpha_i^{-\mathcal{S}_1}(\boldsymbol{h}) - \mathbb{E}\big[\alpha_i^{-\mathcal{S}_1}(\boldsymbol{h}) \big| \boldsymbol{X}_i\big] \Big\}$$
$$+ a_i^{-\mathcal{S}_1}(\boldsymbol{h})^3 n_0^{-1} \sum_{i \in \mathcal{S}_1} \int_0^1 (1 - u) q_3\big(m(\boldsymbol{X}_i) + u a_i^{-\mathcal{S}_1}(\boldsymbol{h})\big) \, du,$$

which appear in an expansion for the cross validation error. Let

$$\begin{aligned}
&\mathcal{E}_1\\
&= \bigcap_{\mathcal{B}:|\mathcal{B}|\le b_n} \bigcap_{1\le k\le K} \bigcap_{i\in\mathcal{S}_k} \bigcap_{d,d'\in\{0\}\cup\mathcal{B}} \Big\{ \sup_{\boldsymbol{h}\in\mathscr{H}_{n,\mathcal{B}}} |\kappa_{d,d'}^{-S_k}(\boldsymbol{X}_i) - \mathbb{E}\kappa_{d,d'}^{-S_k}(\boldsymbol{X}_i)| \le \epsilon_n/(|\mathcal{B}|\log|\mathcal{B}|) \Big\},\\
&\mathcal{E}_2 = \bigcap_{\mathcal{B}:|\mathcal{B}|<b_n} \bigcap_{1\le k\le K} \bigcap_{i\in\mathcal{S}_k} \bigcap_{d\in\mathcal{B}^c} \Big\{ \sup_{\boldsymbol{h}\in\mathscr{H}_{n,\mathcal{B}}} |\hat{\beta}_d^{-S_k}(\boldsymbol{X}_i;\boldsymbol{h})| \le K_0 n^{-1} D^{-1} \Big\},\\
&\mathcal{E}_3 = \bigcap_{\mathcal{B}:|\mathcal{B}|<b_n} \bigcap_{1\le k\le K} \bigcap_{i\in\mathcal{S}_k} \Big\{ \sup_{\boldsymbol{h}\in\mathscr{H}_{n,\mathcal{B}}} |r_{\mathcal{B}}(\boldsymbol{X}_i;\boldsymbol{h}) - \hat{\beta}_0^{-S_k}(\boldsymbol{X}_i;\boldsymbol{h})| \le \epsilon_n \Big\},
\end{aligned}$$

which revise the events $\tilde{\mathcal{E}}_j$, $j = 1, 2, 3$, to accommodate simultaneous prediction targets at $\boldsymbol{x} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. This helps to construct probability bounds for the cross validation error averaged over the K folds. The same techniques used for proving Lemmas 1 and 2 can be applied to construct the following probability bounds:

$$\mathbb{P}(\mathcal{E}_1^c) \le 2b_n^2 D^{b_n} e^{-k_0 a_n \epsilon_n^2/(b_n \log b_n)^2}, \qquad \mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_2^c) \le 2nD^{2+b_n} e^{-k_0 a_n},$$
$$\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_2 \cap \mathcal{E}_3^c) \le 2nD^{b_n} e^{-k_0 a_n \epsilon_n^2/\log n}.$$

Lemma 3. *Suppose that $\bigcap_{0\le j\le 3} \mathcal{E}_j$ holds. Then, for $\mathcal{B} \supset \mathcal{A}$,*

$$\begin{aligned}
&\Big|(Kn_0)^{-1} \sum_{k=1}^{K} \sum_{i\in\mathcal{S}_k} L\big(Y_i, \hat{\beta}_0^{-\mathcal{S}_k}(\boldsymbol{X}_i;\boldsymbol{h})\big) - n^{-1} \sum_{i=1}^{n} L\big(Y_i, m(\boldsymbol{X}_i)\big) - G_1(\boldsymbol{h}) - G_2(\boldsymbol{h})\Big|\\
&\quad \le K_0 \Xi.
\end{aligned}$$

Moreover, $G_1(\boldsymbol{h})$ and $G_2(\boldsymbol{h})$ satisfy, up to an additive error $K_0 \sum_{j=0}^{3} \mathbb{P}(\mathcal{E}_j^c) = o(1)$,

$$\begin{aligned}
\mathbb{P}\big(|G_1(\boldsymbol{h}) - g_1(\boldsymbol{h})| \ge K_0 a_0(t,\boldsymbol{h})\big) &\le K_0 n|\mathcal{B}| e^{-nt\prod_{j\in\mathcal{B}} h_j}, &\text{(S.22)}\\
\mathbb{P}\Big(|G_2(\boldsymbol{h})| \ge K_0\big(\prod_{j\in\mathcal{B}} h_j^{1/2}\big)\big(\omega\bar{h}^2\sqrt{t} + t\big)\Big) &\le K_0 n|\mathcal{B}| e^{-nt\prod_{j\in\mathcal{B}} h_j},
\end{aligned}$$

where $g_1(\boldsymbol{h})$ is a non-random positive function of $\boldsymbol{h}$ with $g_1(\boldsymbol{h}) = O\big(\omega^2\bar{h}^4 + n^{-1}\prod_{j\in\mathcal{B}} h_j^{-1}\big)$, which can be sharpened to $\Omega\big(\omega^2\bar{h}^4 + n^{-1}\prod_{j\in\mathcal{B}} h_j^{-1}\big)$ if the assumption (A2) is strengthened to (A2'), and

$$\begin{aligned}
a_0(t,\boldsymbol{h}) = {}& K_0|\mathcal{B}|^4\omega^2\bar{h}^4 q_n(\log n) \prod_{j\in\mathcal{B}} h_j^{-1/2} + |\mathcal{B}|^2\omega^4\bar{h}^6\\
&+ \sqrt{t}\Big\{K_0\omega\bar{h}^2 q_n(\log n) \prod_{j\in\mathcal{B}} h_j^{-1/2} + |\mathcal{B}|^2\omega^2\bar{h}^4\Big\}\\
&+ t\Big\{|\mathcal{B}|^2\omega\bar{h}^2 + K_0|\mathcal{B}| q_n(\log n) \prod_{j\in\mathcal{B}} h_j^{-1/2}\Big\} + t^{3/2}|\mathcal{B}|^2.
\end{aligned}$$

Lemma 3 expands the cross validation error as a sum of three terms, and provides probability bounds for the last two, $G_1(\boldsymbol{h})$ and $G_2(\boldsymbol{h})$. Note that $G_1(\boldsymbol{h})$ is stochastically bigger than $G_2(\boldsymbol{h})$ with large probability.

Proof of Lemma 3. Assume without loss of generality that $\mathcal{S}_1 = \{1, \ldots, n_0\}$. Note, by Lemma 2, that

$$\Big| n_0^{-1}\sum_{i=1}^{n_0} L\big(Y_i, \hat\beta_0^{-\mathcal{S}_1}(\boldsymbol{X}_i;\boldsymbol{h})\big) - n_0^{-1}\sum_{i=1}^{n_0} L\big(Y_i, m(\boldsymbol{X}_i)\big) - G_1(\boldsymbol{h}) - G_2(\boldsymbol{h})\Big| \le K_0\Xi,$$

which proves the first assertion of the lemma.

Write $G_1(\boldsymbol{h}) = \frac{1}{2}\sum_{j=1}^{4} G_{1,j}(\boldsymbol{h})$, where

$$\begin{aligned}
G_{1,1}(\boldsymbol{h}) &= n_0^{-1}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i) R_0^{*-\mathcal{S}_1}(\boldsymbol{X}_i)^2,\\
G_{1,2}(\boldsymbol{h}) &= n_0^{-1}(1-n_0/n)^{-2}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i)\Big\{\sum_{j\notin\mathcal{S}_1}\nu_{0,j}^{-\mathcal{S}_1}(\boldsymbol{X}_i) w_j\Big\}^2,\\
G_{1,3}(\boldsymbol{h}) &= 2n_0^{-1}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i) R_0^{*-\mathcal{S}_1}(\boldsymbol{X}_i)(1-n_0/n)^{-1}\sum_{j\notin\mathcal{S}_1}\nu_{0,j}^{-\mathcal{S}_1}(\boldsymbol{X}_i) w_j,\\
G_{1,4}(\boldsymbol{h}) &= n_0^{-1}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i) Q^{-\mathcal{S}_1}(\boldsymbol{X}_i)\Big\{Q^{-\mathcal{S}_1}(\boldsymbol{X}_i)\\
&\qquad + 2R_0^{*-\mathcal{S}_1}(\boldsymbol{X}_i) + 2(1-n_0/n)^{-1}\sum_{j\notin\mathcal{S}_1}\nu_{0,j}^{-\mathcal{S}_1}(\boldsymbol{X}_i) w_j\Big\}.
\end{aligned}$$

Note that $G_{1,1}(\boldsymbol{h}), G_{1,2}(\boldsymbol{h}), G_{1,3}(\boldsymbol{h})$ can be expressed as linear combinations of

$$\begin{aligned}
& n_0^{-1}(n-n_0)^{-2}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i)(h_{d_1}h_{d_2}h_{d_1'}h_{d_2'})^{-2}\Big(\prod_{j'\in\mathcal{B}^c} h_{j'}^2\Big)\\
&\times\Big\{\sum_{j\notin\mathcal{S}_1} v(\boldsymbol{X}_j)V_j^{(d_1)}(\boldsymbol{X}_i)V_j^{(d_2)}(\boldsymbol{X}_i)V_j^{(d_1')}(\boldsymbol{X}_i)V_j^{(d_2')}(\boldsymbol{X}_i)K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{X}_j)R_{d_1,d_2,\boldsymbol{X}_j}(\boldsymbol{X}_i)\Big\}^2,
\end{aligned}$$

$$\begin{aligned}
& n_0^{-1}(n-n_0)^{-2}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i)(h_{d_1'}h_{d_2'})^{-2}\Big(\prod_{j'\in\mathcal{B}^c} h_{j'}^2\Big)\\
&\quad\times\Big\{\sum_{j\notin\mathcal{S}_1} V_j^{(d_1')}(\boldsymbol{X}_i)V_j^{(d_2')}(\boldsymbol{X}_i)K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{X}_j)w_j\Big\}^2,
\end{aligned}$$

$$\begin{aligned}
& n_0^{-1}(n-n_0)^{-2}\sum_{i=1}^{n_0} v(\boldsymbol{X}_i)(h_{d_1}h_{d_2}h_{d_1'}h_{d_2'})^{-1}\Big(\prod_{j'\in\mathcal{B}^c} h_{j'}^2\Big)\\
&\quad\times\Big\{\sum_{j\notin\mathcal{S}_1} V_j^{(d_1')}(\boldsymbol{X}_i)K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{X}_j)w_j\Big\}\\
&\quad\times\sum_{\ell\notin\mathcal{S}_1} v(\boldsymbol{X}_\ell)V_\ell^{(d_1)}(\boldsymbol{X}_i)V_\ell^{(d_2)}(\boldsymbol{X}_i)V_\ell^{(d_2')}(\boldsymbol{X}_i)K_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{X}_\ell)R_{d_1,d_2,\boldsymbol{X}_\ell}(\boldsymbol{X}_i),
\end{aligned}$$

respectively, over $d_1', d_2' \in \{0\} \cup \mathcal{B}$ and $d_1, d_2 \in \mathcal{B}$. It follows by Bernstein's inequalities for U-statistics that

$$\mathbb{P}\big(|G_{1,1}(\boldsymbol{h}) - \mathbb{E}G_{1,1}(\boldsymbol{h})| \geq \omega^2 \bar{h}^4 K_0 \sqrt{t}\big) \leq K_0 |\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j},$$

$$\mathbb{P}\Big(|G_{1,2}(\boldsymbol{h}) - \mathbb{E}G_{1,2}(\boldsymbol{h})| \geq n^{-1}\Big(\prod_{j \in \mathcal{B}} h_j^{-1}\Big) K_0 \sqrt{t}\Big) \leq K_0 e^{-nt \prod_{j \in \mathcal{B}} h_j},$$

$$\mathbb{P}\big(|G_{1,3}(\boldsymbol{h})| \geq K_0 \sqrt{t}\big) \leq K_0 |\mathcal{B}| e^{-nt}.$$

Note also that $\mathbb{P}\big(|G_{1,4}(\boldsymbol{h})| \geq K_0 a_0(t, \boldsymbol{h})\big) \leq K_0 n |\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j}$ by Lemma 2, and that

$$\mathbb{P}\Big(|G_2(\boldsymbol{h})| \geq K_0 \Big(\prod_{j \in \mathcal{B}} h_j^{1/2}\Big)\big(\omega \bar{h}^2 \sqrt{t} + t\big)\Big) \leq K_0 n |\mathcal{B}| e^{-nt \prod_{j \in \mathcal{B}} h_j}.$$

Lemma 3 follows by combining the above results, with $g_1(\boldsymbol{h}) = \sum_{j=1}^{3} \mathbb{E}\big[G_{1,j}(\boldsymbol{h})\big] > 0$. ▮

A.2. Proof of Theorem 1

Let $\boldsymbol{a} = (a_1, \ldots, a_D)^\top$ satisfy $a_d = \infty$ for $d \in \mathcal{A}^c$ and $a_d = (n\omega^2)^{-1/(4+|\mathcal{A}|)}$ for $d \in \mathcal{A}$. Define events

$$\mathcal{E}_{4,1} = \big\{G_1(\boldsymbol{a}) + G_2(\boldsymbol{a}) \leq K_0 \omega^2 (n\omega^2)^{-4/(4+|\mathcal{A}|)} + \lambda_n \epsilon_n\big\},$$

$$\mathcal{E}_{4,2} = \bigcap_{\mathcal{B}: |\mathcal{B}| < b_n} \big\{\sup_{\boldsymbol{h} \in \mathscr{H}_{n,\mathcal{B}}} (G_1(\boldsymbol{h}) + G_2(\boldsymbol{h})) \leq -K_0 \lambda_n \epsilon_n\big\},$$

$$\mathcal{E}_{5,1} = \big\{|G_1(\boldsymbol{a}) + G_2(\boldsymbol{a})| \leq K_0 \omega^2 (n\omega^2)^{-4/(4+|\mathcal{A}|)}\big\},$$

$$\mathcal{E}_{5,2} = \Big\{\big|G_1(\hat{\boldsymbol{h}}) + G_2(\hat{\boldsymbol{h}}) - g_1(\hat{\boldsymbol{h}})\big| \leq K_0\big(\omega \|\hat{\boldsymbol{h}}_{\hat{\mathcal{A}}}\|_\infty^2 + n^{-1/2} \prod_{j \in \hat{\mathcal{A}}} \hat{h}_j^{-1/2}\big)^2\Big\}.$$

The four events defined above provide upper bounds for the cross validation errors. Specifically, $\mathcal{E}_{4,1}$ and $\mathcal{E}_{5,1}$ consider different bounds under the special case $\boldsymbol{h} = \boldsymbol{a}$, $\mathcal{E}_{4,2}$ provides a uniform upper bound over $\boldsymbol{h} \in \cup_{|\mathcal{B}| < b_n} \mathscr{H}_{n,\mathcal{B}}$, and $\mathcal{E}_{5,2}$ bounds the deviance of the cross validation error from $g_1(\boldsymbol{h})$ when $\boldsymbol{h} = \hat{\boldsymbol{h}}$. It follows by Lemma 2 that, up to an additive error $K_0 \sum_{j=0}^{3} \mathbb{P}(\mathcal{E}_j^c) = o(1)$,

$$\mathbb{P}(\mathcal{E}_{4,1}^c) \leq K_0 n |\mathcal{A}| \exp\Big(-k_0 \omega^2 (\omega^2 n)^{4/(4+|\mathcal{A}|)} \times \Big\{\frac{\lambda_n \omega (\omega^2 n)^{2/(4+|\mathcal{A}|)} \epsilon_n}{|\mathcal{A}|^3 + K_0 (\log n) |\mathcal{A}| \sqrt{\log(2|\mathcal{A}|)}} \wedge \frac{(\lambda_n \epsilon_n)^{2/3}}{|\mathcal{A}|^{4/3}}\Big\}\Big) \to 0, \tag{S.23}$$

$$\mathbb{P}(\mathcal{E}_{4,2}^c) \leq K_0 n b_n D^{b_n} \exp\Big(-k_0 \Big\{\frac{n \lambda_n^2 \epsilon_n^2}{b_n^2 (\log n)^4} \wedge \big(\lambda_n \epsilon_n \sqrt{n a_n}\big)\Big\}\Big) \to 0, \tag{S.24}$$

$$\mathbb{P}(\mathcal{E}_{5,1}^c) \leq K_0 n |\mathcal{A}| e^{-k_0 \omega^2 (\omega^2 n)^{4/3(4+|\mathcal{A}|)}} \to 0, \tag{S.25}$$

$$\mathbb{P}(\mathcal{E}_{5,2}^c) \leq K_0 n |\hat{\mathcal{A}}| e^{-k_0 (\omega^2 n)^{1/(4+|\hat{\mathcal{A}}|)}} \to 0. \tag{S.26}$$

Define

$$\mathcal{E}_6 = \bigcap_{\mathcal{B}:|\mathcal{B}|\le b_n} \left\{ \sup_{\boldsymbol{h}\in\mathscr{H}_{n,\mathcal{B}}} \left| n^{-1}\sum_{i=1}^n L\big(Y_i, r_{\mathcal{B}}(\boldsymbol{X}_i;\boldsymbol{h})\big) - \mathbb{E}\big[L\big(Y, r_{\mathcal{B}}(\boldsymbol{X};\boldsymbol{h})\big)\big]\right| \le \epsilon_n \right\}.$$

Then it follows by Bernstein's inequality and the definition of $r_{\mathcal{B}}$ that

$$\mathbb{P}(\mathcal{E}_6^c) \le K_0 D^{b_n} e^{-k_0 n \epsilon_n^2}. \tag{S.27}$$

To see that the additive error Ξ has an insignificant order $O(n^{-1})$ when h is sufficiently small, note that for any $h = o\big(\zeta_n^{-1} \wedge 1/\log n\big)$, we have

$$\begin{aligned}&\log\left\{n^{(\alpha\vee 2)+1} e^{-c/h + c[(\alpha\vee 2)-1]\zeta_n}\right\}\\ &\quad = \quad -ch^{-1}\big\{1 - [(\alpha\vee 2)-1]h\zeta_n - [(\alpha\vee 2)+1](h/c)\log n\big\} \to -\infty,\end{aligned}$$

so that

$$n^{\alpha\vee 2} D^{(\alpha\vee 2)-1} e^{-c/h} = O\big(n^{\alpha\vee 2} e^{-c/h + c[(\alpha\vee 2)-1]\zeta_n}\big) = o(n^{-1}).$$

Assume that $\mathcal{E}_0$, $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_6$ hold. By comparing the objective functions (3.2) with $\boldsymbol{h}$ substituted respectively by $\hat{\boldsymbol{h}}$ and $\boldsymbol{a}$, we have

$$\begin{aligned} 0 &\ge n^{-1}\sum_{i=1}^n L\big(Y_i, r_{\mathcal{B}}(\boldsymbol{X}_i;\hat{\boldsymbol{h}})\big) - n^{-1}\sum_{i=1}^n L\big(Y_i, m(\boldsymbol{X}_i)\big) + o(1)\\ &= \mathbb{E}\big[L(Y, r_{\mathcal{B}}(\boldsymbol{X};\hat{\boldsymbol{h}}))\big|\mathscr{D}_n\big] - \mathbb{E}\big[L(Y, m(\boldsymbol{X}))\big] + o(1). \end{aligned} \tag{S.28}$$

By convexity of L and the fact $m(\boldsymbol{X}) = \operatorname{argmin}_{a\in\mathbb{R}}\mathbb{E}\big[L(Y,a)\big|\boldsymbol{X}\big]$, the above inequality implies

$$\int \big\{m(\boldsymbol{x}) - r_{\mathcal{B}}(\boldsymbol{x};\hat{\boldsymbol{h}})\big\}^2 f(\boldsymbol{x})\,d\boldsymbol{x} = o(1),$$

which holds only if $\hat{\mathcal{A}} \supset \mathcal{A}$. It then follows by Lemma 1 and (S.27) that

$$\mathbb{P}(\mathcal{A}\subset\hat{\mathcal{A}}) \ge 1 - K_0\big\{\mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c) + \mathbb{P}(\mathcal{E}_6^c) + \mathbb{P}(\mathcal{E}_0^c)\big\} \to 1.$$

Thus, we have shown that our local linear estimator is consistent if and only if $\hat{\boldsymbol{h}}_{\mathcal{A}} = o_p(1)$. Assume in addition that $\mathcal{E}_1$ holds. Setting $\mathcal{B} = \hat{\mathcal{A}}$ and $\boldsymbol{h} = \hat{\boldsymbol{h}}$ in Lemma 3, and noting (S.28), we have

$$\begin{aligned} &0 \ge -G_1(\boldsymbol{a}) - G_2(\boldsymbol{a}) + O\big(\lambda_n\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)}\big)\\ &+\lambda_n\Big[|\hat{\mathcal{A}}| - |\mathcal{A}| - \sum_{d\in\hat{\mathcal{A}}}\hat{h}_d^4\big\{1+o(1)\big\}\Big] + K_0\Xi + G_2(\hat{\boldsymbol{h}}) + \lambda_n\sum_{d\in\hat{\mathcal{A}}^c}\big\{1+\log(1+\hat{h}_d^4)\big\}^{-1}. \end{aligned} \tag{S.29}$$

The events $\mathcal{E}_{4,1}$ and $\mathcal{E}_{4,2}$ imply that $G_1(\boldsymbol{a}) + G_2(\boldsymbol{a}) = o(\lambda_n)$ and $G_2(\hat{\boldsymbol{h}}) = o(\lambda_n)$, respectively. Thus, noting that $|\hat{\mathcal{A}}| \ge |\mathcal{A}|$ and $\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)} = o\left(\lambda_n\right)$, (S.29) implies that $|\hat{\mathcal{A}}| = |\mathcal{A}|$ on $\mathcal{E}_{4,1} \cap \mathcal{E}_{4,2}$. It then follows by Lemma 1, (S.2), (S.23), (S.24) and (S.27) that

$$\mathbb{P}(\mathcal{A} = \hat{\mathcal{A}}) \geq 1 - K_0\Big\{\sum_{j=0}^{3} \mathbb{P}(\mathcal{E}_j^c) + \mathbb{P}(\mathcal{E}_{4,1}^c) + \mathbb{P}(\mathcal{E}_{4,2}^c) + \mathbb{P}(\mathcal{E}_6^c)\Big\} \to 1,$$

which proves part (i) of Theorem 1.

Next we prove that the empirical bandwidth $\hat{h}_d$ has the asymptotic order stated in parts (ii) or (iii) of Theorem 1, according as $d \in \mathcal{A}$ or $d \notin \mathcal{A}$. Assuming (A2') and using (S.22) and (S.29), we have

$$o\big(\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)}\big) + G_1(\boldsymbol{a}) + G_2(\boldsymbol{a}) \geq G_1(\hat{\boldsymbol{h}}) + G_2(\hat{\boldsymbol{h}}) > 0,$$

so that, on the event $\mathcal{E}_{5,1} \cap \mathcal{E}_{5,2}$,

$$O\big(\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)}\big) \geq \Omega\Big(\omega^2\|\hat{\boldsymbol{h}}_{\mathcal{A}}\|_\infty^4 + n^{-1}\prod_{d\in\mathcal{A}} \hat{h}_d^{-1}\Big) > 0.$$

It then follows by minimality of the order $\omega^2(n\omega^2)^{-4/(4+|\mathcal{A}|)}$, Lemma 1, (S.2) and (S.23)–(S.27) that

$$\begin{aligned}&\mathbb{P}\big(k_0(\omega^2 n)^{-1/(4+|\mathcal{A}|)} \leq \min_{d\in\mathcal{A}} \hat{h}_d \leq \max_{d\in\mathcal{A}} \hat{h}_d \leq K_0(\omega^2 n)^{-1/(4+|\mathcal{A}|)}\big)\\ &\quad\geq 1 - K_0\Big\{\sum_{j=0}^{3} \mathbb{P}(\mathcal{E}_j^c) + \sum_{i=4}^{5}\sum_{j=1}^{2} \mathbb{P}(\mathcal{E}_{i,j}^c) + \mathbb{P}(\mathcal{E}_6^c)\Big\} \to 1, \qquad \text{(S.30)}\end{aligned}$$

which proves part (ii) of Theorem 1.

For $d \in \mathcal{A}^c$, we have, by (S.29), that with probability larger than $1 - K_0\big\{\sum_{j=0}^{3} \mathbb{P}(\mathcal{E}_j^c) + \sum_{i=4}^{5}\sum_{j=1}^{2} \mathbb{P}(\mathcal{E}_{i,j}^c) + \mathbb{P}(\mathcal{E}_6^c)\big\}$,

$$K_0\omega^2(\omega^2 n)^{-4/(4+|\mathcal{A}|)} \geq \sum_{d\in\mathcal{A}^c} \lambda_n\big\{1+\log(1+\hat{h}_d^4)\big\}^{-1} \geq \lambda_n\big\{1+\log(1+\min_{d\in\mathcal{A}^c} \hat{h}_d^4)\big\}^{-1},$$

so that $\log\big(\min_{d\in\mathcal{A}^c} \hat{h}_d^4\big) \geq k_0\lambda_n\omega^{-2}(n\omega^2)^{4/(4+|\mathcal{A}|)}$, which proves part (iii) of Theorem 1 by the same arguments as those used for establishing (S.30). ▌

A.3. Proof of Theorem 2

Part (i) follows from Theorem 1 and Lemma 2, while part (ii) follows from Theorem 1 and Lemma 3.

Note that

$$\begin{aligned}&\Big(n\prod_{d\in\mathcal{A}} \hat{h}_d\Big)^{1/2}\Big\{\hat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) - (\mu_{1,2}/2)\sum_{d\in\mathcal{A}} \hat{h}_d^2 \nabla_{d,d} m(\boldsymbol{x})\Big\}\\ &\quad= \Big(n\prod_{d\in\mathcal{A}} \hat{h}_d\Big)^{1/2}\sum_{i=1}^{n} \nu_{0,i}(\boldsymbol{x})w_i + o_p(1).\end{aligned}$$

To prove part (iii), it suffices to check asymptotic normality of

$$\Big(n\prod_{d\in\mathcal{A}}h_d\Big)^{1/2}\sum_{i=1}^{n}\nu_{0,i}(\boldsymbol{x})w_i \tag{S.31}$$

at $\boldsymbol{h}=\hat{\boldsymbol{h}}$. Let $\phi_{d,d'}(\boldsymbol{x})=\mathbb{E}[\kappa_{d,d'}(\boldsymbol{x})]$ for $d,d'\in\{0\}\cup\mathcal{A}$, and $\big[\breve{\phi}_{d,d'}(\boldsymbol{x})\big]$ be the inverse of the $(|\mathcal{A}|+1)$-dimensional square matrix $\big[\phi_{d,d'}(\boldsymbol{x})\big]$. Thus, $\breve{\kappa}_{d,d'}(\boldsymbol{x})=\breve{\phi}_{d,d'}(\boldsymbol{x})+o_p(1)$. Using Theorem 1(iii), we assume without loss of generality that all inactive variables have been removed from the model and redefine, with slight abuse of notation, $\boldsymbol{h}=\boldsymbol{h}_{\mathcal{A}}$ to be an $|\mathcal{A}|$-dimensional vector. Define $\boldsymbol{h}^*=(h_d^*:d\in\mathcal{A})=\operatorname{argmin}_{\boldsymbol{h}}g_1(\boldsymbol{h})$ and, for any $k>0$, $\mathcal{H}_k=\big\{\boldsymbol{h}\in(0,K_0]^{|\mathcal{A}|}:\max_{d\in\mathcal{A}}|h_d/h_d^*-1|\le k\big\}$. We first treat (S.31) as a random process $\{\mathscr{P}_n(\boldsymbol{h}):\boldsymbol{h}\in\mathcal{H}_{K_0}\}$ and prove its asymptotic Gaussianity. To show finite-dimensional convergence of $\mathscr{P}_n$, consider T arbitrary bandwidth vectors, $\boldsymbol{h}^{(1)},\dots,\boldsymbol{h}^{(T)}$, in $\mathcal{H}_{K_0}$. Denote by $\nu_{0,i}^{(t)}$ the counterpart of $\nu_{0,i}$ with $\boldsymbol{h}$ replaced by $\boldsymbol{h}^{(t)}$. Let $S_{i,t}=(n\prod_{d\in\mathcal{A}}h_d^{(t)})^{1/2}\nu_{0,i}^{(t)}(\boldsymbol{x})w_i$ and $S_i=(S_{i,1}\dots S_{i,T})^\top$. Note that $\mathcal{K}_{\mathcal{A}}^1(\boldsymbol{x};f)=f_{\mathcal{A}}$. Using Theorem 1(ii), it can be shown that $\operatorname{Var}(\sqrt{n}S_i)=\Sigma+o(1)$, for some nonsingular dispersion matrix Σ. We next check the Lyapunov condition. Let $s_{n,1}^2=\sum_{i=1}^n\operatorname{Var}(S_{i,1})$, so that $s_{n,1}^{2+\delta_0}=\Omega(1)$, where δ_0 is specified in (A1). Note that for $d\in\mathcal{A}$,

$$\begin{aligned}
&\mathbb{E}\Big|(h_d^{(1)})^{-1}V_1^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}^{(1)}}(\boldsymbol{X}_1-\boldsymbol{x})w_1\Big|^{2+\delta_0}\\
&\quad=(h_d^{(1)})^{-(2+\delta_0)}\mathbb{E}\Big[\big|V_1^{(d)}(\boldsymbol{x})K_{\boldsymbol{h}^{(1)}}(\boldsymbol{X}_1-\boldsymbol{x})\big|^{2+\delta_0}\mathbb{E}\big[|w_1|^{2+\delta_0}\big|\boldsymbol{X}_1\big]\Big]\\
&\quad=O\Big(\prod_{d\in\mathcal{A}}(h_d^{(1)})^{-(1+\delta_0)}\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}
s_{n,1}^{-(2+\delta_0)}\sum_{i=1}^n\mathbb{E}|S_{i,1}|^{2+\delta_0}&=ns_{n,1}^{-(2+\delta_0)}\mathbb{E}\big|\nu_{0,1}^{(1)}(\boldsymbol{x})w_1\big|^{2+\delta_0}\\
&=O\Big\{n\big(n^{-1}\prod_{d\in\mathcal{A}}h_d^{(1)}\big)^{(2+\delta_0)/2}\big(\prod_{d\in\mathcal{A}}h_d^{(1)}\big)^{-(1+\delta_0)}\Big\}\\
&=O\Big\{\big(n\prod_{d\in\mathcal{A}}h_d^{(1)}\big)^{-\delta_0/2}\Big\}=o(1),
\end{aligned}$$

so that $\sum_{i=1}^n S_i$ is asymptotically Gaussian with mean $\mathbf{0}$ and dispersion matrix Σ by the central limit theorem. For asymptotic Gaussianity of the process $\mathscr{P}_n$, it remains to verify the equicontinuity condition. Note that

$$\mathscr{P}_n(\boldsymbol{h})=\Big(n^{-1}\prod_{d\in\mathcal{A}}h_d\Big)^{1/2}\breve{\phi}_{0,0}(\boldsymbol{x})\sum_{i=1}^nK_{\boldsymbol{h}}(\boldsymbol{X}_i-\boldsymbol{x})w_i+o_p(1).$$

Define, for $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}_{K_0}$,

$$\mathcal{F}(\boldsymbol{h}, \boldsymbol{h}') = n^{-1/2} \sum_{i=1}^{n} \Big\{ \Big(\prod_{d \in \mathcal{A}} h_d \Big)^{1/2} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) - \Big(\prod_{d \in \mathcal{A}} h'_d \Big)^{1/2} K_{\boldsymbol{h}'}(\boldsymbol{X}_i - \boldsymbol{x}) \Big\} w_i$$

and, for $\delta > 0$, $\boldsymbol{Z}(\delta) = \sup \big\{ |\mathcal{F}(\boldsymbol{h}, \boldsymbol{h}')| : \boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}_{K_0},\ \max_{d \in \mathcal{A}} h_d^{*-1} |h_d - h'_d| \leq \delta \big\}$. The equicontinuity condition holds when $\boldsymbol{Z}(\delta)$ is sufficiently small with high probability. Assume that the kernel K is supported within $[-K_0, K_0]$. Then we have, for a Rademacher sequence $\{E_i\}$ independent of $\mathscr{D}_n$ and setting $h_o = (\omega n)^{-1/(4+|\mathcal{A}|)}$, that

$$\begin{aligned}
&\mathbb{E}\boldsymbol{Z}(\delta) \\
&\leq 2\, \mathbb{E} \sup \Big\{ \Big| n^{-1/2} \sum_{i : \|\boldsymbol{X}_i - \boldsymbol{x}\|_\infty \leq K_0 h_o} w_i E_i \{ (\prod_{d \in \mathcal{A}} h_d)^{1/2} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) - (\prod_{d \in \mathcal{A}} h'_d)^{1/2} K_{\boldsymbol{h}'}(\boldsymbol{X}_i - \boldsymbol{x}) \} \Big| : \\
&\qquad\qquad\qquad \boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}_{K_0},\ \max_{d \in \mathcal{A}} h_d^{*-1} |h_d - h'_d| \leq \delta \Big\} \\
&\leq 2 n^{-1/2} h_o^{-|\mathcal{A}|/2} \delta K_0 \mathbb{E} \sup \Big\{ \Big| \sum_{i : \|\boldsymbol{X}_i - \boldsymbol{x}\|_\infty \leq K_0 h_o} w_i E_i \Big| : \boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}_{K_0},\ \max_{d \in \mathcal{A}} h_d^{*-1} |h_d - h'_d| \leq \delta \Big\} \\
&\leq \delta K_0.
\end{aligned}$$

Applying the concentration theorem (Bousquet, 2002), we have, for $z > 0$,

$$\mathbb{P}\Big(\boldsymbol{Z}(\delta) \geq \delta K_0 + z K_0 \big(1 + \delta n h_o^{|\mathcal{A}|} \log n\big)^{1/2} + z^2 K_0 (n h_o^{|\mathcal{A}|})^{1/2} \log n \Big) \leq \exp(-n z^2),$$

which proves the equicontinuity condition. This, together with its finite-dimensional asymptotic Gaussianity, implies that $\mathscr{P}_n$ converges weakly to a zero-mean Gaussian process $\mathbb{G}$. That $\mathbb{G}(\boldsymbol{h})$ has a constant variance $\mu_{2,0}^{|\mathcal{A}|} \sigma(\boldsymbol{x})^2 f_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})^{-1} v(\boldsymbol{x})^{-2}$ follows from the fact that for any $\boldsymbol{h} \in \mathcal{H}_{K_0}$,

$$\begin{aligned}
&\mathrm{Var}\Big((n^{-1} \prod_{d \in \mathcal{A}} h_d)^{1/2} \breve{\phi}_{0,0}(\boldsymbol{x}) \sum_{i=1}^{n} K_{\boldsymbol{h}}(\boldsymbol{X}_i - \boldsymbol{x}) w_i \Big) \\
&\quad = \mu_{2,0}^{|\mathcal{A}|} \sigma(\boldsymbol{x})^2 f_{\mathcal{A}}(\boldsymbol{x}_{\mathcal{A}})^{-1} v(\boldsymbol{x})^{-2} + o(1).
\end{aligned}$$

Part (iii) then follows by Slutsky's theorem and in-probability convergence of $\hat{h}_d / h_d^*$ to a deterministic limit for each $d \in \mathcal{A}$. ∎

Funding

Supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 17307819, 17307321).

References

[1] Allen, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *J. Comput. Graph. Statist.* **22**, 284–299. MR3173715

[2] Belloni, A. & Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39**, 82–130. MR2797841

[3] Bertin, K. & Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.* **2**, 1224–1241. MR2461900

[4] Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334**, 495–500. MR1890640

[5] Chen, J., Zhang, C., Kosorok, M. R. & Liu, Y. (2017). Double sparsity kernel learning with automatic variable selection and data extraction. arXiv:1706.01426. MR3858117

[6] Fan, J., Feng, Y. & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557. MR2847969

[7] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall: London. MR1383587

[8] Fan, J., Heckman, N. E. & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141–150. MR1325121

[9] Fan, J., Hu, T. C. & Truong, Y. K. (1994). Robust non-parametric function estimation. *Scand. J. Statist.* **21**, 433–446. MR1310087

[10] Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **70**, 849–911. MR2530322

[11] Giordano, F., Lahiri, S. N. & Parrella, M. L. (2020). GRID: A variable selection and structure discovery method for high dimensional nonparametric regression. *Ann. Statist.* **48**, 1848–1874. MR4124346

[12] Giordano, F. & Parrella, M. L. (2016). Bias-corrected inference for multivariate nonparametric regression: model selection and oracle property. *J. Multivariate Anal.* **143**, 71–93. MR3431420

[13] Huang, J., Horowitz, J. L. & Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282–2313. MR2676890

[14] Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *J. Roy. Statist. Soc. Ser. B* **60**, 271–293. MR1616041

[15] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101. MR0161415

[16] Lafferty, J., & Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.* **36**, 28–63. MR2387963

[17] Ledoux, M. & Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag: Berlin. MR1102015

[18] Li, R., Zhong, W. & Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129–1139. MR3010900

[19] Li, Y., & Zhu, J. (2008). L1-norm quantile regression. *J. Comput. Graph. Statist.* **17**, 163–185. MR2424800

[20] Lin, C. Y., Bondell, H., Zhang, H. H. & Zou, H. (2013). Variable se-

lection for non-parametric quantile regression via smoothing spline analysis of variance. *Stat* **2**, 255–268. MR4027316

[21] Lin, Y. & Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–2297. MR2291500

[22] Meier, L., Van de Geer, S. & Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37**, 3779–3821. MR2572443

[23] Miller, H. & Hall, P. (2010). Local polynomial regression and variable selection. In *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, Vol. 6, 216–233. IMS Collections. MR2798521

[24] Park,, M. Y. & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**, 659–677. MR2370074

[25] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7**, 186–199. MR1128411

[26] Radchenko, P. & James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Amer. Statist. Assoc.* **105**, 1541–1553. MR2796570

[27] Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B* **71**, 1009–1030. MR2750255

[28] Rosasco, L., Villa, S., Mosci, S., Santoro, M. & Verri, A. (2013). Nonparametric sparsity and regularization. *J. Mach. Learn. Res.* **14**, 1665–1714. MR3104492

[29] Stefanski, L. A., Wu, Y., & White, K. (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *J. Amer. Statist. Assoc.* **109**, 574–589. MR3223734

[30] Storlie, C. B., Bondell, H. D., Reich, B. J. & Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica* **21**, 679–705. MR2829851

[31] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505–563. MR1419006

[32] van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645. MR2396809

[33] van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag: New York. MR1385671

[34] White, K. R., Stefanski, L. A. & Wu, Y. (2017). Variable selection in kernel regression using measurement error selection likelihoods. *J. Amer. Statist. Assoc.* **112**, 1587–1597. MR3750883

[35] Wu, Y. & Stefanski, L. A. (2015). Automatic structure recovery for additive models. *Biometrika* **102**, 381–395. MR3371011

[36] Yang, L., Lv, S. & Wang, J. (2016). Model-free variable selection in reproducing kernel Hilbert space. *J. Mach. Learn. Res.* **17**, 1–24. MR3517105

[37] Yang, Y., & Tokdar, S. T. (2015). Minimax-optimal nonparametric

regression in high dimensions. *Ann. Statist.* **43**, 652–674. MR3319139

[38] Ye, G. B. & Xie, X. (2012). Learning sparse gradients for variable selection and dimension reduction. *J. Mach. Learn. Res.* **87**, 303–355. MR2917060

[39] Zhang, H. H. & Lin, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statist. Sinica* **16**, 1021–1041. MR2281313