

Shape constrained kernel-weighted least squares: Estimating production functions for Chilean manufacturing industries

This appendix includes:

- Extensions to SCKLS and a description of the relationship between SCKLS, CNLS and CWB (Appendix A),
- Technical proofs of the theoretical results (Appendix B).
- A test of affinity based on SCKLS (Appendix C)
- An algorithm for SCKLS computational performance (Appendix D).
- Comprehensive results of existing and additional numerical experiments (Appendix E).
- Description of a semiparametric partially linear model to integrate contextual variable (Appendix F).
- Details about the application to the Chilean manufacturing data (Appendix G)

A More on SCKLS, CNLS and CWB

In this section, we first give details on the extensions and practical considerations to SCKLS. We then mention some recently proposed estimators that are related to SCKLS, and make connections and comparisons among these methods.

A.1 More on practical considerations and extensions to SCKLS

A.1.1 SCKLS with general constraints

We focus on global concavity/convexity and monotonicity constraints in the main manuscript. But the SCKLS estimator can handle any types of shape constrained by imposing constraints on decision variables $\{a_i, \mathbf{b}_i\}_{i=1}^m$. We re-define the SCKLS estimator as

$$\begin{aligned} \min_{\mathbf{a}, \mathbf{b}} \quad & \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 K \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}} \right) \\ \text{subject to} \quad & l(\mathbf{x}_i) \leq \hat{g}^{(\mathbf{s})}(\mathbf{x}_i | \mathbf{a}, \mathbf{b}) \leq u(\mathbf{x}_i), \quad i = 1, \dots, m \end{aligned} \tag{A.1}$$

where $\mathbf{a} = (a_1, \dots, a_m)'$ and $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_m)'$. $l(\cdot)$ and $u(\cdot)$ represent lower and upper bounds at each evaluation point respectively. \mathbf{s} denotes the order of partial derivative to each evaluation point \mathbf{x}_i .

A.1.2 SCKLS with Local Polynomial

With the proposed estimator in (A.1), we are only able to impose the constraints by using the functional estimate and/or first partial derivatives. For constraints involving a higher order of derivatives, we need to formulate SCKLS estimator with a higher order local polynomial function. For the multivariate local polynomial, we borrow the following

notation from Masry (1996).

$$\begin{aligned}
\mathbf{r} &= (r_1, \dots, r_d), & \mathbf{r}! &= r_1! \times \dots \times r_d!, & \bar{\mathbf{r}} &= \sum_{k=1}^d r_k, \\
\mathbf{x}^{\mathbf{r}} &= x_1^{r_1} \times \dots \times x_d^{r_d}, & \sum_{0 \leq \bar{\mathbf{r}} \leq p} &= \sum_{k=0}^p \sum_{r_1=0}^k \dots \sum_{r_d=0}^k, & & \text{and} \\
(D^{\mathbf{r}} g)(\mathbf{x}) &= \frac{\partial^{\mathbf{r}} g(\mathbf{x})}{\partial x_1^{r_1} \dots \partial x_d^{r_d}}
\end{aligned}$$

With this notation, we can approximate any function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ locally (around any \mathbf{x}) using a multivariate polynomial of total order p , given by

$$g(\mathbf{z}) := \sum_{0 \leq \bar{\mathbf{r}} \leq p} \frac{1}{\mathbf{r}!} (D^{\bar{\mathbf{r}}} g)(\mathbf{x}) (\mathbf{z} - \mathbf{x})^{\bar{\mathbf{r}}}. \quad (\text{A.2})$$

We now define the SCKLS estimator with a local polynomial function of order p as follows:

$$\begin{aligned}
\min_{\mathbf{b}_i} \quad & \sum_{i=1}^m \sum_{j=1}^n \left(y_j - \sum_{0 \leq \bar{\mathbf{r}} \leq p} \mathbf{b}'_i (\mathbf{X}_j - \mathbf{x}_i)^{\bar{\mathbf{r}}} \right)^2 K \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}} \right) \\
\text{subject to} \quad & l(\mathbf{x}_i) \leq \hat{g}^{(s)}(\mathbf{x}_i | \mathbf{b}) \leq u(\mathbf{x}_i), \quad i = 1, \dots, m
\end{aligned} \quad (\text{A.3})$$

where \mathbf{b}_i is the functional or derivative estimates at each evaluation points and $\mathbf{b} = (\mathbf{b}'_1, \dots, \mathbf{b}'_m)'$. When we select $p = 1$, then the problem becomes exactly same as the proposed estimator in (A.1). This extension allows us to make the proposed methods more general and applicable for other applications of shape restricted functional estimation in which higher order derivative restricts may be required. From a computational complexity point of view, it is still optimizing a quadratic objective function within a convex solution space, and thus, the problem is still typically solvable within polynomial time.

As demonstrated in Li and Racine (2007), the rate of convergence of local polynomial estimator is the same for $p = 1$ and $p = 2$. From a theoretical perspective, one could attempt to select a polynomial estimator with $p \geq 3$ to improve its convergence performance (at least theoretical). But that would require much stronger assumption on the smoothness of g_0 , and would lead to additional computational burden¹. Our experience suggests that

¹While the optimization problem is still polynomial time solvable, the number of decision variables

SCKLS inherits these properties from the local polynomial method. Therefore, in practice, with only monotonicity and concavity/convexity constraints, we feel that it suffices to consider SCKLS with $p = 1$ (i.e. local linear).

A.1.3 SCKLS with k -nearest neighbor

Our primary application of interest is production functions estimated for census manufacturing data where the input distributions are often highly skewed meaning there are many small establishments, but relatively few large establishments². To address this issue, we propose to use a k -nearest neighbor (k -NN) approach in SCKLS which we will refer to as SCKLS k -NN which is in spirit similar to the extension to the CWB-type estimator proposed by Li et al. (2016). The k -NN approach uses a smaller bandwidth for smoothing in dense data regions and a larger bandwidth when the data is sparse. For a further description of the method, see for example Li and Racine (2007). For any given k , the formulation of SCKLS k -NN with monotonicity and concavity constraints leads to a different weighting scheme in the objective function, as illustrated in the following.

$$\begin{aligned}
& \min_{a_i, \mathbf{b}_i} \quad \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 w \left(\frac{\|\mathbf{X}_j - \mathbf{x}_i\|}{R_{\mathbf{x}_i}} \right) \\
& \text{subject to} \quad a_i - a_l \geq \mathbf{b}_i'(\mathbf{x}_i - \mathbf{x}_l), \quad i, l = 1, \dots, m \\
& \quad \quad \quad \mathbf{b}_i \geq 0, \quad i = 1, \dots, m
\end{aligned} \tag{A.4}$$

where $w(\cdot)$ is a general weight function, $\|\cdot\|$ is the Euclidean norm and $R_{\mathbf{x}_i}$ denotes the Euclidean distance between \mathbf{x}_i and k -th nearest neighbor of \mathbf{x}_i among the set of all covariates $\{\mathbf{X}_j\}_{j=1}^n$. In practice, k can be chosen by leave-one-out cross validation (LOOCV).

would increase and the constraint matrix would become significantly more dense, lending to computational challenges.

²An establishment is defined as a single physical location where business is conducted or where services or industrial operations are performed.

A.1.4 SCKLS with non-uniform grid

As noted in the paper, the SCKLS estimator requires the user to specify the number and locations of the evaluation points. We can also address the input skewness issue by constructing the evaluation points differently, using a non-uniform grid method. To do so, we first use kernel density estimation to estimate the density function for each input dimension. Then we take the equally spaced percentiles of the estimated density function and construct non-uniform grid. Figure A.1 demonstrates how the non-uniform grid are constructed for the 2-dimensional case. In this example, we set the minimum and maximum of the observed inputs (with respect to each coordinate) as the edge of the grid, and compute equally spaced percentile. When the support of the covariates is non-regular (e.g. not a hyperrectangle), we shall limit ourselves to evaluation points inside the convex hull of $\{\mathbf{X}_j\}_{j=1}^n$.

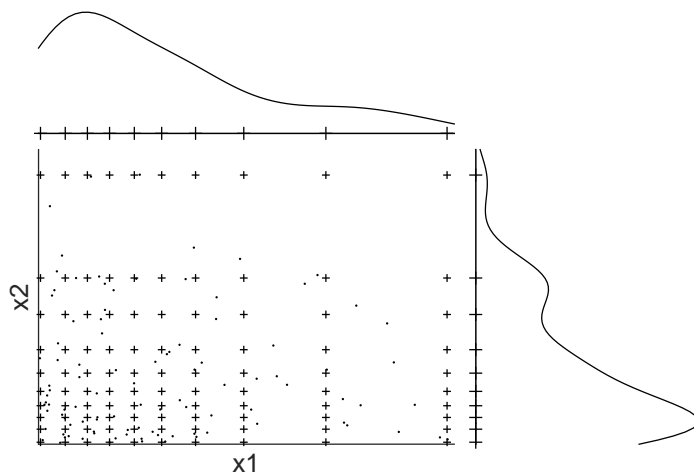


Figure A.1. Example of non-uniform grid with kernel density estimation.

A.2 Some related work

A.2.1 Convex Nonparametric Least Squares (CNLS)

Kuosmanen (2008) extends Hildreth’s least squares approach to the multivariate setting with a multivariate input vector, and coins the term “Convex Nonparametric Least Squares” (CNLS)³. CNLS builds upon the assumption that the true but unknown production function g_0 belongs to the class of monotonically increasing and globally concave functions, denoted by G_2 in this paper. Given the observations $\{\mathbf{X}_j, y_j\}_{j=1}^n$, a set of unique fitted values, $\hat{y}_j = \hat{\alpha}_j + \hat{\beta}_j' \mathbf{X}_j$, can be found by solving the quadratic programming (QP) problem

$$\begin{aligned} \min_{\alpha, \beta} \quad & \sum_{j=1}^n (y_j - (\alpha_j + \beta_j' \mathbf{X}_j))^2 \\ \text{subject to} \quad & \alpha_j + \beta_j' \mathbf{X}_j \leq \alpha_l + \beta_l' \mathbf{X}_j, \quad j, l = 1, \dots, n \\ & \beta_j \geq 0, \quad j = 1, \dots, n \end{aligned} \tag{A.5}$$

where α_j and β_j define the intercept and slope parameters that characterize the estimated set of hyperplanes. The inequality constraints in (A.5) can be interpreted as a system of Afriat inequalities (Afriat, 1972; Varian, 1984) to impose concavity constraints. We emphasize that CNLS does not assume or restrict the domain G_2 to only piece-wise affine functions. We also note that the functional estimates resulting from (A.5) is unique only at the observed data points. In addition, when $d = 1$, Chen and Wellner (2016) and Ghosal and Sen (2016) proved that the CNLS-type estimator attains $n^{-1/2}$ pointwise rate of convergence if the true function is piece-wise linear.

Finally, we remark that CNLS is related to the method of sieves (Grenander, 1981; Chen and Qiu, 2016) in the following way. The estimator could be rewritten as

$$\hat{g}_n \in \operatorname{argmin}_{g \in \mathcal{G}^n} \frac{1}{n} \sum_{j=1}^n (y_j - g(\mathbf{X}_j))^2,$$

where $\mathcal{G}^n = \{g : \mathbb{R}^d \rightarrow \mathbb{R} \mid g(\mathbf{x}) = \min_{j \in \{1, \dots, n\}} (\alpha_j + \beta_j' \mathbf{x}), \text{ with } \beta_j \geq 0 \text{ for } j = 1, \dots, n\}$.

³A related maximum likelihood formulation was proposed by Banker and Maindiratta (1992), with its consistency proved by Sarath and Maindiratta (1997).

However, since the sets $\mathcal{G}^1, \mathcal{G}^2, \dots$ are not compact, most known results on sieves do not directly apply here.

A.2.2 Constrained Weighted Bootstrap (CWB)

A.2.2.1 Introduction

Hall and Huang (2001) proposed the monotone kernel regression method in univariate function. Du et al. (2013) generalized this model to handle multiple general shape constraints for multivariate functions, which they refer to as Constrained Weighted Bootstrap (CWB). CWB estimator is constructed by introducing weights for each observed data point. The weights are selected to minimize the distance to unconstrained estimator while satisfying the shape constraints. The function is estimated as

$$\hat{g}(\mathbf{x}|\mathbf{p}) = \sum_{j=1}^n p_j A_j(\mathbf{x}) y_j \quad (\text{A.6})$$

where $\mathbf{p} = (p_1, \dots, p_n)'$, p_j is the weights introduced for each observation and $A_j(\mathbf{x})$ is a local weighting matrix (e.g. local linear kernel weighting matrix). Du et al. (2013) relaxed the restriction imposed by Hall and Huang (2001) that p_j is non-negative and propose to calculate \mathbf{p} by minimizing its distance to unrestricted weights, $\mathbf{p}_u = (1/n, \dots, 1/n)'$, under derivative-based shape constraints⁴. The problem is formulated as follows.

$$\begin{aligned} \min_{\mathbf{p}} \quad & D(\mathbf{p}) = \sum_{j=1}^n (p_j - p_u)^2 = \sum_{j=1}^n (p_j - 1/n)^2 \\ \text{subject to} \quad & l(\mathbf{x}_i) \leq \hat{g}^{(\mathbf{s})}(\mathbf{x}_i|\mathbf{p}) \leq u(\mathbf{x}_i), \quad i = 1, \dots, m \end{aligned} \quad (\text{A.7})$$

where \mathbf{x}_i represents a set of points for evaluating constraints, the elements of \mathbf{s} represent the order of partial derivative, and $g^{\mathbf{s}}(\mathbf{x}) = [\partial^{s_1} g(\mathbf{x}) \cdots \partial^{s_r} g(\mathbf{x})] / [\partial x_1^{s_1} \cdots \partial x_r^{s_r}]$ for $\mathbf{s} = (s_1, s_2, \dots, s_r)$. Here the shape restrictions (e.g. concavity/convexity and monotonicity constraints) are imposed at a set of evaluation points $\{\mathbf{x}_i\}_{i=1}^m$ through setting appropriate lower and upper bounds to the corresponding partial derivatives of the function. One way to

⁴The use of the equality constraint $\sum_j p_j = 1$ in Du et al. (2013) is a typo, and this condition is not used by them. In fact, it may harm the estimation procedure. Our empirical results show that this equality constraint only makes difference in very few cases and the difference is typically small.

interpret the CWB estimator is as a two-step process: 1) estimate an unconstrained kernel estimator; 2) find the shape constrained function that is as close as possible (as measured by the Euclidean distance in p -space) to the unconstrained kernel estimator. Based on our experience, CWB tends to suffer from computational difficulties and occasionally poor estimates in small samples. We suggest changing the objective function to minimize the distance from the estimated function to the observed data. This modification seems to improve the estimates empirically as shown in Appendix E.

A.2.2.2 CWB estimator that minimize the distance from the observed data

We propose an extension of the CWB estimator by converting the objective function from p -space to y -space. Instead of minimizing the distance between the unconstrained estimator and the shape restricted functional estimate by minimizing the distance between the two functions in p -space, we propose to minimize the distance between the observed vector of \mathbf{y} and the shape restricted functional estimates in y -space. The estimator, which we shall refer to as CWB in y -space, is formulated as follows:

$$\begin{aligned} \min_{\mathbf{p}} \quad & D_y(\mathbf{p}) = \sum_{j=1}^n (y_j - \hat{g}(\mathbf{X}_j|\mathbf{p}))^2 \\ \text{subject to} \quad & l(\mathbf{x}_i) \leq \hat{g}^{(s)}(\mathbf{x}_i|\mathbf{p}) \leq u(\mathbf{x}_i), \quad i = 1, \dots, m, \\ & \sum_{j=1}^n p_j = 1. \end{aligned} \tag{A.8}$$

Since the objective function is not necessarily convex in \mathbf{p} , this problem is a general nonlinear optimization problem which is harder to solve.

A.2.2.3 Calculating the first partial derivative of $\hat{g}(\mathbf{x}|\mathbf{p})$ for CWB

Du et al. (2013) proposed the CWB estimator which requires estimating the first partial derivatives of unconstrained functional estimates, $\hat{g}^{(1)}(\mathbf{x}|\mathbf{p})$. Here, we test two different methods of calculating the partial derivatives. The first method is to calculate the numerical derivative, $\hat{g}^{(1)}(\mathbf{x}|\mathbf{p}) = \frac{\hat{g}(\mathbf{x}+\Delta|\mathbf{p}) - \hat{g}(\mathbf{x}|\mathbf{p})}{\Delta}$, to obtain the approximated derivative estimate. Racine (2016) shows that the numerical derivative is very close to the analytic derivative.

The second method is to use the slope estimates of local linear estimator directly as a proxy for the first partial derivative. We evaluate the performance of CWB in p -space estimator with these two different methods. Table A.1 and Table A.2 summarize the RMSE performance against the true function on the observed points and the evaluation points respectively. The experimental setting is based on Experiment 1 in Section 5.

Table A.1. RMSE on observation points for different methods to obtain $\hat{g}^{(1)}(\mathbf{x}|\mathbf{p})$.

Number of observations		Average RMSE on the observation points				
		100	200	300	400	500
2-input	Numerical derivative	0.260	0.163	0.143	0.153	0.164
	Slope estimates of LL	0.421	0.357	0.284	0.306	0.293
3-input	Numerical derivative	0.236	0.256	0.208	0.246	0.240
	Slope estimates of LL	0.356	0.427	0.336	0.294	0.279
4-input	Numerical derivative	0.259	0.226	0.222	0.216	0.210
	Slope estimates of LL	0.388	0.397	0.276	0.261	0.259

Table A.2. RMSE on evaluation points for different methods to obtain $\hat{g}^{(1)}(\mathbf{x}|\mathbf{p})$.

Number of observations		Average RMSE on the evaluation points				
		100	200	300	400	500
2-input	Numerical derivative	0.284	0.188	0.157	0.176	0.193
	Slope estimates of LL	0.445	0.387	0.321	0.334	0.323
3-input	Numerical derivative	0.309	0.355	0.272	0.331	0.271
	Slope estimates of LL	0.438	0.507	0.403	0.371	0.363
4-input	Numerical derivative	0.408	0.381	0.354	0.333	0.308
	Slope estimates of LL	0.530	0.535	0.396	0.387	0.368

The results show that CWB using the numerical derivative performs better than CWB using the slope estimates from the local linear kernel estimator particularly when the sample size is small.

A.3 A comparison between SCKLS, CNLS and CWB

Figure A.2 is meant to be illustrative of the relationship between the SCKLS, CNLS and CWB estimators in a two-dimensional estimated ϵ -space where there are more than two observations, but for the rest of the $n - 2$ observations, their estimated ϵ_j s are held fix. The

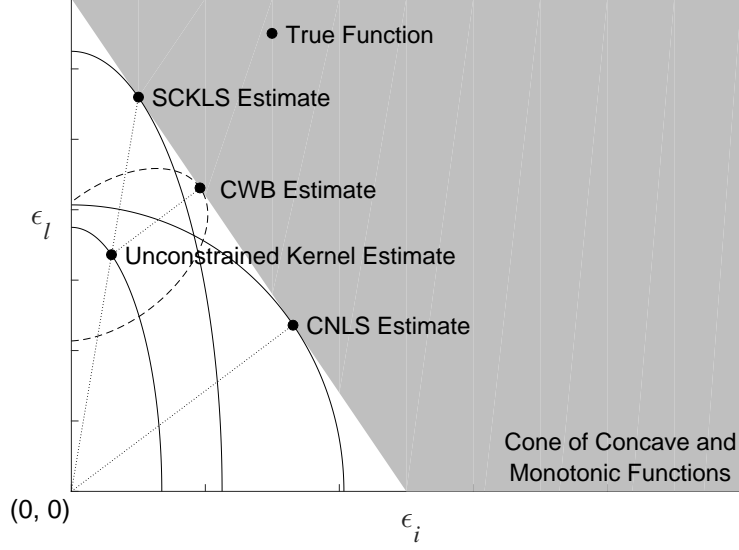


Figure A.2. Comparison of different estimators in the estimated- ϵ -space.

gray area indicates the cone of concave and monotonic functions. CNLS estimates a monotonic and concave function while minimizing the sum of squared errors, that is, minimizing the distance from the origin to the cone in the estimated ϵ -space. CWB estimates a monotonic and concave function by finding the closest point, measured in p -space, on the cone of concave and monotonic functions to unconstrained kernel estimate. SCKLS minimizes a weighted function of estimated errors, and therefore avoids overfitting the observed data. However, as shown in B.2, SCKLS can be interpreted as minimizing the weighted distance from the unconstrained local linear kernel estimator to the cone of concave and monotonic functions.

A.3.1 CNLS as a Special Cases of SCKLS

Let \hat{g}_n and \hat{g}_n^{CNLS} denote the SCKLS estimator and the CNLS estimator respectively. We will next examine the relationship between them.

Assumption A.1. *The set of evaluation points is equal to the set of sample input vectors, i.e. $m = n$ and $\mathbf{x}_i = \mathbf{X}_i$ for $i = 1, \dots, n$.*

Proposition A.1. *Suppose that Assumption A.1 holds. Then, for any n , when the vector*

of bandwidth goes to zero, i.e. $\|\mathbf{h}\| \rightarrow \mathbf{0}$ (where $\mathbf{h} = (h_1, \dots, h_d)'$), the SCKLS estimator \hat{g}_n converges to the CNLS estimator \hat{g}_n^{CNLS} pointwise at $\mathbf{X}_1, \dots, \mathbf{X}_n$.

Proposition A.1 essentially says that CNLS can be viewed as a special case of SCKLS. Note that in comparison to the CNLS estimator, our SCKLS estimator has tuning parameters, which to some extent control the bias–variance tradeoff (in a non-trivial way given the shape restrictions). For reasonable values of these tuning parameters, SCKLS estimator performs better than CNLS. See also Section 5 of the main manuscript. This is especially true for the estimates close to the boundary of the input space, where imposing the shape constraint alone could lead to severe overfitting of the data, and thus biased estimates. Indeed, in view of Theorem 3 (from the main manuscript), we have that $\sup_{\mathcal{S}} |\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})| = o_p(1)$, while on the other hand, $\sup_{\mathcal{S}} |\hat{g}_n^{CNLS}(\mathbf{x}) - g_0(\mathbf{x})|$ does not converge to zero in probability.

Additional equivalence results can also be shown. Proposition A.2 shows the equivalence of linear regression subject to monotonicity constraints and the SCKLS estimator when the bandwidth vector approaches infinity.

Proposition A.2. *Given Assumption 1(v). For any given n , when the bandwidth vector goes to infinity (i.e. $\min_{k=1, \dots, d} h_k \rightarrow \infty$), the SCKLS estimator converges to the least squares estimator of the linear regression model subject to monotonicity constraints.*

A.3.2 CWB in y -space as a Special Cases of SCKLS

Let \hat{g}_n and $\hat{g}_n^{CWB Y}$ denote the SCKLS estimator and the CWB y -space estimator respectively. We will next examine the relationship between them.

Proposition A.3. *Suppose that Assumption A.1 holds. Then, for any n , when the vector of bandwidth goes to zero for both the SCKLS estimator and the CWB in y -space estimator, i.e. $\|\mathbf{h}\| \rightarrow \mathbf{0}$ (where $\mathbf{h} = (h_1, \dots, h_d)'$), the SCKLS estimator \hat{g}_n converges to the CWB in y -space estimator $\hat{g}_n^{CWB Y}$ pointwise at $\mathbf{X}_1, \dots, \mathbf{X}_n$.*

Proposition A.3 states that SCKLS and CWB in y -space estimators converge to the same estimates as $\|\mathbf{h}\| \rightarrow \mathbf{0}$. Combining with Proposition A.1, CNLS can be viewed as a special case of SCKLS and CWB in y -space.

A.3.3 The relationship between CWB in p -space and SCKLS

Again start from the SCKLS estimator, and in view of Assumption 1 (v), for any sufficiently small \mathbf{h} , we have

$$K\left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}}\right) = \begin{cases} 0 & \text{if } \mathbf{x}_i \neq \mathbf{X}_j, \\ K(\mathbf{0}) & \text{if } \mathbf{x}_i = \mathbf{X}_j, \end{cases} \text{ for } \forall i, j.$$

Then, the objective function of the SCKLS estimator (3) is equal to $\sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0})$, and thus

$$\operatorname{argmin}_{a_1, b_1, \dots, a_n, b_n} \sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0}) = \operatorname{argmin}_{a_1, \dots, a_n} \sum_{j=1}^n (y_j - a_j)^2 = \operatorname{argmin}_{a_1, \dots, a_n} L(g(a_j))$$

where $L(\cdot) = \sum_{j=1}^n (\cdot)^2$ is the squared error loss function, $g(a_j) = y_j - a_j$ the definition of the residual.

Alternatively now consider the objective function of CWB, specifically $D(\mathbf{p}) = \sum_{j=1}^n (p_u - p_j)^2 = \sum_{j=1}^n (1/n - p_j)^2 = L(m(g(p_j)))$. And let $L(\cdot)$ continue to be defined as above as the squared error lost function and $g(p_j)$ as the definition of the residual. This implies that $m(\cdot) = \frac{\cdot}{y_j n}$. Therefore, the CWB estimator can be interpreted as a projection of a local polynomial estimator to the cone of functions which are monotonic and concave in which the direction of projection minimizes a specific weighting of the unconstrained local polynomial residuals in which the weights are defined as $\frac{1}{y_j n}$. Therefore, even if the vector of bandwidth goes to zero for the CWB in p -space estimator, i.e. $\|\mathbf{h}\| \rightarrow \mathbf{0}$ (where $\mathbf{h} = (h_1, \dots, h_d)'$), the CWB estimator and CNLS are not equivalent because the y_j in the denominator of the weights is not a function of the bandwidth.

A.3.4 On the computational aspects

We also compare the computational burden of each estimators. Table A.3 shows the size of quadratic programming problems of each estimators: SCKLS, CNLS and CWB. The size of a quadratic programming problem of the SCKLS estimator is fully controllable because the number of decision variables and constraints is a function of the number of

evaluation points and independent of the number of observed points. Because of this, we can solve large-scale problems with $n > 100,000$ using the SCKLS estimator while other shape constrained nonparametric estimators might face prohibitive computational difficulties without any data pre-processing.

Table A.3. The size of quadratic programming problems of each estimator.

	SCKLS	CNLS	CWB
Number of decision variables	$m(d+1)$	$n(d+1)$	n
Number of global concavity constraints	$m(m-1)$	$n(n-1)$	$m(m-1)$

B Technical proofs

B.1 Summary of the proof strategy

Theorems 1–4 concern the consistency and convergence rate of the SCKLS estimator and serve as the primary results in our theoretical development. As such, before presenting the technical details, we summarize our proof strategy as follows:

1. We rewrite the SCKLS estimator, after some manipulations, as the projection of the local linear estimator to a convex cone of monotonic and concave functions under a certain norm. More precisely, the SCKLS estimator

$$\hat{g}_n \in \operatorname{argmin}_{g \in G_2} \|g - \tilde{g}_n\|_{n,m}^2,$$

where \tilde{g}_n is the local linear estimator, G_2 is the set that contains all the concave and increasing functions, and $\|\cdot\|_{n,m}$ is a norm defined in detail later in Appendix B.2.

2. (Theorem 1). Let \hat{g}_n be the SCKLS estimator and $g_0 \in G_2$ be the truth. Using the new formulation of SCKLS above, we see that

$$\|\hat{g}_n - \tilde{g}_n\|_{n,m} \leq \|g_0 - \tilde{g}_n\|_{n,m}.$$

Moreover, by the triangular inequality, we have that

$$\|\hat{g}_n - g_0\|_{n,m} \leq \|\hat{g}_n - \tilde{g}_n\|_{n,m} + \|\tilde{g}_n - g_0\|_{n,m} \leq 2\|\tilde{g}_n - g_0\|_{n,m}.$$

Using the results on the uniform consistency of the local linear estimator (e.g. Fan and Guerre (2016), see our Lemma B.1 and Lemma B.2), we can bound the RHS of the triangle inequality equation by $O_p(n^{-2/(4+d)} \log n) = o_p(1)$. Consequently, $\|\hat{g}_n - g_0\|_{n,m}$ converges to zero at the same rate. To complete the proof, we show that the discrete L_2 distance between \hat{g}_n and g_0 is bounded above by a constant times $\|\hat{g}_n - g_0\|_{n,m}$.

3. (Theorem 2). Building upon Theorem 1, we then make use of the concavity of \hat{g}_n and g_0 to establish uniform consistency. Loosely speaking, this relies on the fact that the convergence in L_2 for a sequence of Lipschitz (and concave) functions implies the uniform convergence in the interior of the domain. See Lemma B.3 and Lemma B.4 below for more detail. Note that we only look at \hat{g}_n on the a compact subset interior of its domain, in order to make sure that \hat{g}_n is Lipschitz there. That is also why we do not have consistency on the boundary from the current proof strategy.
4. (Theorem 3). If we let the number of evaluation points, m , grow at a certain rate slower than n , we can extend the uniform consistency result to the entire support of \mathbf{X} . The assumption on the rate of growth of m makes sure that the first partial derivative of SCKLS, $\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x})$, is bounded for some positive constant, so the SCKLS is Lipschitz over the entire domain.
5. (Theorem 4). This can be viewed as a generalization of Theorem 2. The main ingredient of its proof is to establish $\|\hat{g}_n - g_0^*\|_{n,m} = o_p(1)$. Then the uniform consistency follows from the concavity of \hat{g}_n and g_0^* via Lemma B.4.

B.2 Alternative definition of SCKLS

Recall that given observations $\{\mathbf{X}_j, y_j\}_{j=1}^n$ and evaluation points $\{\mathbf{x}_i\}_{i=1}^m$, the (unconstrained) local linear estimator at \mathbf{x}_i is $(\tilde{a}_i, \tilde{\mathbf{b}}_i)$ for $i = 1, \dots, m$, where $(\tilde{a}_1, \tilde{\mathbf{b}}_1, \dots, \tilde{a}_m, \tilde{\mathbf{b}}_m)$

is the (unique) minimizer of

$$\sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 K \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}} \right).$$

For simplicity, we assume that the bandwidth is equal for all input dimensions, i.e. $\mathbf{h} = (h, \dots, h)'$. Since the objective function is quadratic, for any $(a_1, \mathbf{b}_1, \dots, a_m, \mathbf{b}_m)$, its value equals

$$nh^d \sum_{i=1}^m (\tilde{a}_i - a_i, (\tilde{\mathbf{b}}_i - \mathbf{b}_i)' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - a_i \\ (\tilde{\mathbf{b}}_i - \mathbf{b}_i) h \end{pmatrix} + \text{Const}$$

where

$$\Sigma_i = \frac{1}{nh^d} \sum_{j=1}^n U \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{h} \right) \left\{ U \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{h} \right) \right\}' K \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{h} \right)$$

with $U(\mathbf{x})$ being the vector $(1, \mathbf{x}')'$ and

$$\text{Const} = \sum_{i=1}^m \sum_{j=1}^n (y_j - \tilde{a}_i - (\mathbf{X}_j - \mathbf{x}_i)' \tilde{\mathbf{b}}_i)^2 K \left(\frac{\mathbf{X}_j - \mathbf{x}_i}{h} \right).$$

Therefore, SCKLS can be simply viewed as a minimizer of

$$\sum_{i=1}^m (\tilde{a}_i - a_i, (\tilde{\mathbf{b}}_i - \mathbf{b}_i)' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - a_i \\ (\tilde{\mathbf{b}}_i - \mathbf{b}_i) h \end{pmatrix}$$

subject to the shape constraints imposed on $(a_1, \mathbf{b}_1, \dots, a_m, \mathbf{b}_m)$. More generally, fixing $\{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ and h , and define a new squared distance measure between two functions g_1, g_2 as

$$\|g_1 - g_2\|_{n,m}^2 = \frac{1}{m} \sum_{i=1}^m \left(g_1(\mathbf{x}_i) - g_2(\mathbf{x}_i), \left(\frac{\partial g_1}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_2}{\partial \mathbf{x}}(\mathbf{x}_i) \right)' h \right) \Sigma_i \begin{pmatrix} g_1(\mathbf{x}_i) - g_2(\mathbf{x}_i) \\ \left(\frac{\partial g_1}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_2}{\partial \mathbf{x}}(\mathbf{x}_i) \right)' h \end{pmatrix},$$

then SCKLS belongs to⁵

$$\operatorname{argmin}_{g \in G_2} \|g - \tilde{g}_n\|_{n,m}$$

where G_2 is the set that contains all the concave and increasing functions from \mathbf{S} to \mathbb{R} .

Below, we list some useful results on the behaviors of Σ_i and $(\tilde{a}_i, \tilde{\mathbf{b}}_i)$. These results follow from Fan and Guerre (2016).

Lemma B.1 (Lemma 5 of Fan and Guerre (2016), Page 508). *Suppose that Assumption 1(i)-1(vi) hold, then with probability one, there exists $C > 1$ such that the eigenvalues of Σ_i are in $[1/C, C]$ for all $i = 1, \dots, m$ for sufficiently large n .*

Lemma B.2 (Proposition 7 of Fan and Guerre (2016), Page 509). *Suppose that Assumption 1(i)-1(vi) hold, then as $n \rightarrow \infty$,*

$$\sup_{i=1, \dots, m} \left(|\tilde{a}_i - g_0(\mathbf{x}_i)|^2, \left\| h \left\{ \tilde{\mathbf{b}}_i - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i) \right\} \right\|^2 \right) = O_p(n^{-4/(4+d)} \log n).$$

B.3 Proof of Theorems in Section 3

B.3.1 Proof of Theorem 1

Proof. With a sufficiently large n , the uniqueness of the estimates of $\hat{g}_n(\mathbf{x}_i)$ and $\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i)$ for $i = 1, \dots, m$ is established because our objective function corresponds to is a quadratic programming problem with a positive definite (strictly convex) objective function with a feasible solution. See Bertsekas (1995).

Based on our characterization of SCKLS in Appendix B.2, we note that the objective function at the SCKLS estimate is smaller than or equal to that at the truth, and thus

$$\|\hat{g}_n - \tilde{g}_n\|_{n,m}^2 \leq \|g_0 - \tilde{g}_n\|_{n,m}^2.$$

⁵To be more precise technically, if $g_1 - g_2$ is not differentiable, then $\|g_1 - g_2\|_{n,m}$ needs to be taken as the infimum among all possible sub-gradients in the previous definition. Nevertheless, since we only consider the behavior of the functions at finitely many points, without loss of generality, here we can restrict ourselves to differentiable functions.

Moreover, by the triangular inequality, we have that

$$\|\hat{g}_n - g_0\|_{n,m} \leq \|\hat{g}_n - \tilde{g}_n\|_{n,m} + \|\tilde{g}_n - g_0\|_{n,m} \leq 2\|\tilde{g}_n - g_0\|_{n,m}.$$

As such,

$$\|\hat{g}_n - g_0\|_{n,m}^2 \leq 4\|\tilde{g}_n - g_0\|_{n,m}^2. \quad (\text{B.1})$$

Recall that the (unconstrained) local linear estimator at \mathbf{x}_i is $(\tilde{a}_i, \tilde{\mathbf{b}}_i)$ for $i = 1, \dots, m$. It follows from Lemma B.2 that

$$\|\tilde{g}_n - g_0\|_{n,m}^2 = \frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0(\mathbf{x}_i) \\ (\tilde{\mathbf{b}}_i - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} = O_p(n^{-4/(4+d)} \log n)$$

In addition, from Lemma B.1, we have that

$$\begin{aligned} \|\hat{g}_n - g_0\|_{n,m}^2 &= \frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i), (\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i) \\ (\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \\ &\geq \frac{1}{Cm} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2, \end{aligned} \quad (\text{B.2})$$

where C is the constant mentioned in the statement of Lemma B.1.

Plugging the above two equations into (B.1) yields

$$\frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 \leq O_p(n^{-4/(4+d)} \log n) = o_p(1).$$

□

B.3.2 Proof of Theorem 2

For the sake of clarity, we have divided the proof of Theorem 2 into several parts.

B.3.2.1 Some useful lemmas

Here we list two useful lemmas on the convergence of convex functions.

Lemma B.3. *Suppose that $f_0, f_1, f_2, \dots : \mathbf{C}' \rightarrow \mathbb{R}$ are Lipschitz and convex functions, where $\mathbf{C}' \subset \mathbb{R}^d$ is a compact and convex set. In addition, assume that these functions all have the same bound and Lipschitz constant. Then*

$$\lim_{n \rightarrow \infty} \int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 d\mathbf{x} = 0$$

implies that

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| = 0$$

for any compact \mathbf{C} in the interior of \mathbf{C}' .

Proof. Suppose that the common Lipschitz constant is $M > 0$. Moreover, suppose that

$$\sup_{\mathbf{x} \in \mathbf{C}'} \inf_{\mathbf{y} \in \mathbf{C}} \|\mathbf{x} - \mathbf{y}\| =: \delta.$$

Essentially, that means that for any $\mathbf{x} \in \mathbf{C}'$, the ball of radius δ centered at \mathbf{x} (denoted as $B_\delta(\mathbf{x})$) intersects with \mathbf{C} .

Next, suppose that $\sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| \geq \epsilon$ for some $\epsilon > 0$. Let

$$\mathbf{x}^* \in \operatorname{argmax}_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})|.$$

Then for any \mathbf{x} that lies inside the ball of radius $\min\{\delta, \epsilon/(4M)\}$ centered at \mathbf{x}^* , we have that

$$\begin{aligned} |f_n(\mathbf{x}) - f_0(\mathbf{x})| &= |f_n(\mathbf{x}) - f_n(\mathbf{x}^*) + f_n(\mathbf{x}^*) - f_0(\mathbf{x}^*) + f_0(\mathbf{x}^*) - f_0(\mathbf{x})| \\ &\geq |f_n(\mathbf{x}^*) - f_0(\mathbf{x}^*)| - |f_n(\mathbf{x}) - f_n(\mathbf{x}^*)| - |f_0(\mathbf{x}^*) - f_0(\mathbf{x})| \\ &\geq \epsilon - \frac{\epsilon}{4M}M - \frac{\epsilon}{4M}M = \frac{\epsilon}{2}, \end{aligned}$$

where we made use of the Lipschitz constant for f_n and f_0 in the second last line above. Consequently,

$$\int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 d\mathbf{x} \geq \left(\frac{\epsilon}{2}\right)^2 \operatorname{Vol}(B_{\min\{\delta, \epsilon/(4M)\}}(\mathbf{x}^*)) = \operatorname{Const.} \times \epsilon^{d+2}$$

for any $0 < \epsilon < 4M\delta$.

But since $\epsilon > 0$ is arbitrary, $\limsup_{n \rightarrow \infty} \sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| \geq \epsilon$ for any sufficiently small ϵ would imply

$$\limsup_{n \rightarrow \infty} \int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 d\mathbf{x} \geq \text{Const.} \times \epsilon^{d+2},$$

violating

$$\lim_{n \rightarrow \infty} \int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 d\mathbf{x} = 0.$$

Our proof is thus completed by contradiction. □

The following Lemma B.4 can be viewed as a small extension of Lemma B.3. This is the version that we shall use in the proof of Theorem 2.

Lemma B.4. *Suppose that $f_0, f_1, f_2, \dots : \mathbf{C}' \rightarrow \mathbb{R}$ are Lipschitz and convex functions (that could be random), where $\mathbf{C}' \subset \mathbb{R}^d$ is a compact and convex set. In addition, assume that these functions all have the same bound and Lipschitz constant. Furthermore, $q : \mathbf{C}' \rightarrow \mathbb{R}$ with $\inf_{\mathbf{x} \in \mathbf{C}'} q(\mathbf{x}) > 0$. Then, for any fixed compact set \mathbf{C} in the interior of \mathbf{C}' ,*

$$\int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 q(\mathbf{x}) d\mathbf{x} \xrightarrow{p} 0$$

implies that

$$\sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| \xrightarrow{p} 0$$

as $n \rightarrow \infty$.

Proof. Following the arguments in the proof of Lemma B.3, we see that $\sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| \geq \epsilon$ would entail

$$\int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 q(\mathbf{x}) d\mathbf{x} \geq \left(\frac{\epsilon}{2}\right)^2 \text{Vol}(B_{\min\{\delta, \epsilon/(4M)\}}(\mathbf{x}^*)) \inf_{\mathbf{x} \in \mathbf{C}} q(\mathbf{x}) = \text{Const.} \times \epsilon^{d+2}$$

for any sufficiently small ϵ . Consequently, $\int_{\mathbf{C}'} \{f_n(\mathbf{x}) - f_0(\mathbf{x})\}^2 q(\mathbf{x}) d\mathbf{x} \xrightarrow{p} 0$ implies that $\sup_{\mathbf{x} \in \mathbf{C}} |f_n(\mathbf{x}) - f_0(\mathbf{x})| \xrightarrow{p} 0$. □

B.3.2.2 Lipschitz continuity of SCKLS

For the reasons that will become clearer later, it is useful to investigate the Lipschitz continuity of SCKLS before we present our proof of Theorem 2. Our finding is summarized in the following lemma. Its proof is similar to that of Proposition 4 of Lim and Glynn (2012, Page 201–202), or that of Theorem 1 of Chen and Samworth (2016, online supplementary material, Page 2–6). We provide a concise version of the proof for the sake of completeness. To better illustrate its main idea and intuition, below we focus on the scenario of $d = 1$.

Lemma B.5. *Under the assumptions of the first part of Theorem 2 (in the case where m increases with n), for any convex and compact set $\mathbf{C} \subset \text{int}(\mathbf{S})$ (where $\text{int}(\cdot)$ denotes the interior of a set), there exists some constants $B > 0$ and $M > 0$ such that \hat{g}_n is B -bounded and M -Lipschitz over \mathbf{C} with probability one as $n \rightarrow \infty$.*

Proof. As explained before, here we focus on the scenario of $d = 1$. Without loss of generality, we can take $\mathbf{S} = [0, 1]$ and $\mathbf{C} = [\delta, 1 - \delta]$ for some $\delta \in (0, 1/2)$.

Let $B_0 = \sup_{[0,1]} |g_0(x)|$. First, we show that the event

$$\sup_{x \in [\delta, 1-\delta]} |\hat{g}_n(x)| \leq 2B_0 + 1 =: B$$

happens with probability one as $n \rightarrow \infty$.

Since \hat{g}_n is increasing, $\sup_{x \in [\delta, 1-\delta]} |\hat{g}_n(x)| = \max(|\hat{g}_n(\delta)|, |\hat{g}_n(1 - \delta)|)$. In addition, due to the monotonicity of \hat{g}_n , suppose that $\hat{g}_n(\delta) \leq 0$, then $|\hat{g}_n(x)| \geq |\hat{g}_n(\delta)|$ for $x \in [0, \delta]$; otherwise, if $\hat{g}_n(\delta) > 0$, $|\hat{g}_n(x)| \geq |\hat{g}_n(\delta)|$ for $x \in [\delta, 2\delta]$ (actually, this statement is true for $x \in [\delta, 1]$; but for our purpose, it suffices to only consider $x \in [\delta, 2\delta]$). As such,

$|\hat{g}_n(\delta)| > 2B_0 + 1$ would imply that

$$\begin{aligned}
\frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 &\geq \frac{\mathbf{1}_{\{\hat{g}_n(\delta) \leq 0\}}}{m} \sum_{i=1}^m (\hat{g}_n(x_i) - g_0(x_i))^2 \mathbf{1}_{\{x_i \in [0, \delta]\}} \\
&\quad + \frac{\mathbf{1}_{\{\hat{g}_n(\delta) > 0\}}}{m} \sum_{i=1}^m (\hat{g}_n(x_i) - g_0(x_i))^2 \mathbf{1}_{\{x_i \in [\delta, 2\delta]\}} \\
&\geq (2B_0 + 1 - B_0)^2 \left(\frac{\mathbf{1}_{\{\hat{g}_n(\delta) \leq 0\}}}{m} \sum_{i=1}^m \mathbf{1}_{\{x_i \in [0, \delta]\}} + \frac{\mathbf{1}_{\{\hat{g}_n(\delta) > 0\}}}{m} \sum_{i=1}^m \mathbf{1}_{\{x_i \in [\delta, 2\delta]\}} \right) \\
&\geq (B_0 + 1)^2 \min \left(\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{x_i \in [0, \delta]\}}, \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{x_i \in [\delta, 2\delta]\}} \right) \\
&\stackrel{n \rightarrow \infty}{\geq} B_0^2 \delta \min_{[0,1]} q(x) > 0.
\end{aligned}$$

where $q(\cdot)$ is the density function with respect to what the empirical distribution of $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ converges to (see Assumption 2(i)). Here the last line also follows from Assumption 2(i). Note that Theorem 1 says that $\frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 = o_p(1)$, which would result in a contradiction. Therefore, $|\hat{g}_n(\delta)| \leq 2B_0 + 1$.

Furthermore, we can reapply the above argument to show that $|\hat{g}_n(1 - \delta)| \leq 2B_0 + 1$. Consequently,

$$\sup_{x \in [\delta, 1-\delta]} |\hat{g}_n(x)| \leq 2B_0 + 1 = B$$

happens with probability one as $n \rightarrow \infty$.

Second, note that the above proof works for any $\delta \in (0, 1/2)$. Therefore, we also have that

$$\sup_{x \in [\delta/2, 1-\delta/2]} |\hat{g}_n(x)| \leq 2B_0 + 1$$

with probability one as $n \rightarrow \infty$.

Finally, since \hat{g}_n is concave, we note that the Lipschitz constant over $[\delta, 1 - \delta]$ is bounded above by

$$\max \left(\frac{|\hat{g}_n(\delta/2) - \hat{g}_n(\delta)|}{\delta/2}, \frac{|\hat{g}_n(1 - \delta/2) - \hat{g}_n(1 - \delta)|}{\delta/2} \right) \leq 4(2B_0 + 1)/\delta =: M.$$

In other words, intuitively speaking, in terms of the Lipschitz constant, the most extreme case for concave functions always occurs on the boundary. For general cases (i.e. $d > 1$),

see for instance, van der Vaart and Wellner (1996, Page 165, Problem 7). \square

B.3.2.3 Putting things together to prove Theorem 2

Proof.

First claim: when m increases with n .

Let C' be a compact and convex set such that $C \subset \text{int}(C')$ and $C' \subset \text{int}(S)$, where $\text{int}(\cdot)$ denotes the interior of a set.

By Lemma B.5, we have that \hat{g}_n is B -bounded and M -Lipschitz over C' with probability one as $n \rightarrow \infty$. Therefore, $\{\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})\}^2 \mathbf{1}_{\{\mathbf{x} \in C'\}}$ belongs to the class of functions that is bounded and equicontinuous over C' . By Theorem 3.1 of (Rao, 1962, Page 662) (which can also be viewed as a generalization of the Uniform Law of Large Numbers; see also Chapter 2.4 of van der Vaart and Wellner (1996)), we have that

$$\left| \frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 \mathbf{1}_{\{\mathbf{x}_i \in C'\}} - \int_{C'} \{\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})\}^2 q(\mathbf{x}) d\mathbf{x} \right| \xrightarrow{p} 0.$$

In addition, it follows from Theorem 1 that

$$o_p(1) = \frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 \geq \frac{1}{m} \sum_{i=1}^m (\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i))^2 \mathbf{1}_{\{\mathbf{x}_i \in C'\}}.$$

Combining the above two equations together yields

$$\int_{C'} \{\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})\}^2 q(\mathbf{x}) d\mathbf{x} = o_p(1).$$

It then follows immediately from Lemma B.4 that as $n \rightarrow \infty$,

$$\sup_{\mathbf{x} \in C} |\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})| \xrightarrow{p} 0.$$

Second claim: when m is fixed.

In views of Lemma B.1 and Theorem 1,

$$\frac{1}{C} \sum_{i=1}^m \left[|\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i)|^2 + \left\| \left(\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i) \right) h \right\|^2 \right] \leq \|\hat{g}_n - g_0\|_{n,m}^2 = O_p(n^{-4/(4+d)} \log n)$$

where the first inequality is from Lemma B.1, and the last equality is from Theorem 1.

Since m is fixed and $h = O(n^{-1/(4+d)})$, it follows from that $|\hat{g}_n(\mathbf{x}_i) - g_0(\mathbf{x}_i)| = O_p(n^{-2/(4+d)} \log n) \xrightarrow{p} 0$ and $\left\| \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i) \right\| = O_p(n^{-1/(4+d)} \log n) \xrightarrow{p} 0$ for every $i = 1, \dots, m$. \square

B.3.3 Proof of Theorem 3

Proof. Using Equation (B.2) but focusing on the difference between the derivatives instead, we have that

$$\frac{h^2}{Cm} \sum_{i=1}^m \left\| \left(\frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i) \right) \right\|^2 \leq \|\hat{g}_n - g_0\|_{n,m}^2 = O_p(n^{-4/(4+d)} \log n)$$

as $n \rightarrow \infty$. It then follows from $h = O(n^{-1/(4+d)})$ and Assumption 3 that

$$\sum_{i=1}^m \left\| \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) - \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}_i) \right\|^2 = O_p(h^{-2} m n^{-4/(4+d)} \log n) = o_p(1).$$

This implies that $\max_{i=1, \dots, m} \left\| \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) \right\|_{\infty} \leq \sup_{\mathbf{x} \in \mathcal{S}} \left\| \frac{\partial g_0}{\partial \mathbf{x}}(\mathbf{x}) \right\|_{\infty} + o_p(1)$. Now since

$$\hat{g}_n(\mathbf{x}) = \min_{i \in \{1, \dots, m\}} \left\{ \hat{g}_n(\mathbf{x}_i) + (\mathbf{x} - \mathbf{x}_i)' \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) \right\},$$

we have that with probability one,

$$\sup_{\mathbf{x} \in \mathcal{S}} \left\| \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}) \right\|_{\infty} \leq M$$

for some $M > 0$, as $n \rightarrow \infty$.

For any $\epsilon > 0$, we can always find a compact set $\mathcal{C}_{\epsilon} \subset \mathcal{S}$ such that $\sup_{\mathbf{x} \in \mathcal{S}} \inf_{\mathbf{y} \in \mathcal{C}_{\epsilon}} \|\mathbf{x} - \mathbf{y}\| < \frac{\epsilon}{2(M+M_{g_0})}$, where M_{g_0} is the Lipschitz constant of g_0 . In view of Theorem 2, $\sup_{\mathbf{x} \in \mathcal{C}_{\epsilon}} |\hat{g}_n(\mathbf{x}) -$

$g_0(\mathbf{x})| \rightarrow 0$ in probability. Therefore,

$$\sup_{\mathbf{x} \in \mathcal{S}} |\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})| \leq \sup_{\mathbf{x} \in \mathcal{C}_\epsilon} |\hat{g}_n(\mathbf{x}) - g_0(\mathbf{x})| + (M + M_{g_0}) \left\{ \sup_{\mathbf{x} \in \mathcal{S}} \inf_{\mathbf{y} \in \mathcal{C}_\epsilon} \|\mathbf{x} - \mathbf{y}\| \right\} \leq \epsilon$$

as $n \rightarrow \infty$. Since ϵ is picked arbitrarily, we have shown the consistency of \hat{g}_n over \mathcal{S} . \square

B.4 Proof of Theorems in Section 4

B.4.1 Proof of Theorem 4

Proof. Using the definition of SCKLS in Appendix B.2 and the notation in the proofs of Theorem 1 and Theorem 2, we have that

$$\begin{aligned} & \sum_{i=1}^m \left(\tilde{a}_i - g_0^*(\mathbf{x}_i), \left(\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)' h \right) \right) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0^*(\mathbf{x}_i) \\ \left(\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)' h \right) \end{pmatrix} \\ & \geq \sum_{i=1}^m \left(\tilde{a}_i - \hat{a}_i, \left(\tilde{\mathbf{b}}_i - \hat{\mathbf{b}}_i' h \right) \right) \Sigma_i \begin{pmatrix} \tilde{a}_i - \hat{a}_i \\ \left(\tilde{\mathbf{b}}_i - \hat{\mathbf{b}}_i' h \right) \end{pmatrix} \\ & = \sum_{i=1}^m \left(\tilde{a}_i - g_0^*(\mathbf{x}_i), \left(\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)' h \right) \right) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0^*(\mathbf{x}_i) \\ \left(\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)' h \right) \end{pmatrix} \\ & \quad + 2 \sum_{i=1}^m \left(\tilde{a}_i - g_0^*(\mathbf{x}_i), \left(\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)' h \right) \right) \Sigma_i \begin{pmatrix} g_0^*(\mathbf{x}_i) - \hat{a}_i \\ \left(\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i' h \right) \end{pmatrix} \\ & \quad + \sum_{i=1}^m \left(g_0^*(\mathbf{x}_i) - \hat{a}_i, \left(\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i' h \right) \right) \Sigma_i \begin{pmatrix} g_0^*(\mathbf{x}_i) - \hat{a}_i \\ \left(\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i' h \right) \end{pmatrix} \end{aligned}$$

where we recall that \hat{a}_i and $\hat{\mathbf{b}}_i$ are respectively the estimated value and its gradient from SCKLS at evaluation point \mathbf{x}_i , i.e., $\hat{a}_i = \hat{g}_n(\mathbf{x}_i)$ and $\hat{\mathbf{b}}_i = \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i)$.

Therefore, in view of Lemma B.2, with probability one, for sufficiently large n ,

$$\frac{2}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \quad (\text{B.3})$$

$$\geq \frac{1}{m} \sum_{i=1}^m (g_0^*(\mathbf{x}_i) - \hat{a}_i, (\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i)' h) \Sigma_i \begin{pmatrix} g_0^*(\mathbf{x}_i) - \hat{a}_i \\ (\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i) h \end{pmatrix} \geq \frac{1}{mC} \sum_{i=1}^m (g_0^*(\mathbf{x}_i) - \hat{a}_i)^2 \quad (\text{B.4})$$

Next, we show that the quantity in (B.3) converges to zero in probability as $n \rightarrow \infty$. The proof can be divided into six steps:

1. The contribution to (B.3) from evaluation points lying outside a carefully pre-chosen compact subset \mathbf{S}' of the interior of \mathbf{S} (denoted as $\text{int}(\mathbf{S})$) can be made arbitrarily small. This follows from the Cauchy–Schwarz inequality that

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x} \notin \mathbf{S}'\}} \\ & \leq \sqrt{\frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0^*(\mathbf{x}_i) \\ (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x} \notin \mathbf{S}'\}}} \end{aligned} \quad (\text{B.5})$$

$$\times \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{a}_i - g_0^*(\mathbf{x}_i), (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix}}. \quad (\text{B.6})$$

Because of Lemma B.1 and Assumption 2(i), the quantity in (B.5) can be made arbitrarily small by choosing \mathbf{S}' sufficiently close to \mathbf{S} . In addition, applying the Cauchy–

Schwarz inequality to (B.3) and comparing it to (B.4) yields

$$\begin{aligned}
& 2 \sqrt{\frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0^*(\mathbf{x}_i) \\ (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix}} \\
& \quad \times \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{a}_i - g_0^*(\mathbf{x}_i), (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix}} \\
& \geq \frac{1}{m} \sum_{i=1}^m (g_0^*(\mathbf{x}_i) - \hat{a}_i, (\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i)' h) \Sigma_i \begin{pmatrix} g_0^*(\mathbf{x}_i) - \hat{a}_i \\ (\frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i) - \hat{\mathbf{b}}_i) h \end{pmatrix},
\end{aligned}$$

so (B.6) is no greater than

$$\begin{aligned}
& 2 \sqrt{\frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \tilde{a}_i - g_0^*(\mathbf{x}_i) \\ (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix}} \\
& \rightarrow 2 \left\{ \int_{\mathcal{S}} (g_0(\mathbf{x}) - g_0^*(\mathbf{x}))^2 Q(d\mathbf{x}) \right\}^{1/2} \leq 2 \left\{ \int_{\mathcal{S}} g_0^2(\mathbf{x}) Q(d\mathbf{x}) \right\}^{1/2}.
\end{aligned}$$

Consequently, the claim in this step is proved.

2. We now investigate the contribution to (B.3) from evaluation points lying inside \mathcal{S}' . Using Lemma B.5, we have that \hat{g}_n is bounded (i.e. from both below and above) and M -Lipschitz over \mathcal{S}' in probability.

Combining this with Lemma B.1 implies that

$$\begin{aligned}
& \left| \frac{1}{m} \sum_{i=1}^m (\tilde{a}_i - g_0^*(\mathbf{x}_i), (\tilde{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i))' h) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x} \in \mathcal{S}'\}} \right. \\
& \quad \left. - \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\mathbf{x}_i), \left(\frac{\partial (g_0 - g_0^*)}{\partial \mathbf{x}}(\mathbf{x}_i) \right)' h \right) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x} \in \mathcal{S}'\}} \right| \rightarrow 0
\end{aligned}$$

in probability. As such, we can instead work on

$$\frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\mathbf{x}_i), \left(\frac{\partial(g_0 - g_0^*)}{\partial \mathbf{x}}(\mathbf{x}_i) \right)' h \right) \Sigma_i \begin{pmatrix} \hat{a}_i - g_0^*(\mathbf{x}_i) \\ (\hat{\mathbf{b}}_i - \frac{\partial g_0^*}{\partial \mathbf{x}}(\mathbf{x}_i)) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x} \in \mathbf{S}'\}} \quad (\text{B.7})$$

3. Next, we bound (and eliminate) the influence from the parts involving partial derivatives of g_0 , g_0^* and \hat{g}_n in (B.7). Since \hat{g}_n is bounded and M -Lipschitz over \mathbf{S}' in probability, together with Lemma B.2, we could bound (B.7) from above by

$$\frac{1}{m} \sum_{i=1}^m (g_0(\mathbf{x}_i) - g_0^*(\mathbf{x}_i))(\hat{g}_n(\mathbf{x}_i)) - g_0^*(\mathbf{x}_i) \mathbf{1}_{\{\mathbf{x} \in \mathbf{S}'\}} + O(h) + O(h^2),$$

which is arbitrarily close to $\frac{1}{m} \sum_{i=1}^m (g_0(\mathbf{x}_i) - g_0^*(\mathbf{x}_i))(\hat{g}_n(\mathbf{x}_i) - g_0^*(\mathbf{x}_i)) \mathbf{1}_{\{\mathbf{x} \in \mathbf{S}'\}}$ as $n \rightarrow \infty$ (i.e. $h \rightarrow 0$). Here we also used the fact that $\sup_{i=1, \dots, m} |\Sigma_i^{(11)} - 1| \rightarrow 0$, where $\Sigma_i^{(11)}$ is the first diagonal entry of the matrix Σ_i .

4. Now we re-expand \hat{g}_n from \mathbf{S}' to \mathbf{S} as

$$\hat{g}_n^{\mathbf{S}'}(\mathbf{x}) = \min_{i \in \{1, \dots, m | \mathbf{x}_i \in \mathbf{S}'\}} \left\{ \hat{g}_n(\mathbf{x}_i) + (\mathbf{x} - \mathbf{x}_i)' \frac{\partial \hat{g}_n}{\partial \mathbf{x}}(\mathbf{x}_i) \right\}.$$

Three useful facts about $\hat{g}_n^{\mathbf{S}'}$ are listed below:

- $\hat{g}_n^{\mathbf{S}'} \geq \hat{g}_n$, with $\hat{g}_n^{\mathbf{S}'}(\mathbf{x}_i) = \hat{g}_n(\mathbf{x}_i)$ for any $\mathbf{x}_i \in \mathbf{S}'$.
- there exists some $B > 0$ such that $\sup_{\mathbf{x} \in \mathbf{S}} \hat{g}_n^{\mathbf{S}'}(\mathbf{x}) \leq B$ in probability. Importantly, given that there is a common compact and convex set \mathbf{C} such that $\mathbf{C} \subset \mathbf{S}'$ for all the \mathbf{S}' to be considered, the constant B does not depend on the choice of \mathbf{S}' . To see this, we note that $\hat{g}_n^{\mathbf{C}} = \hat{g}_n$ over \mathbf{C} , which is also B' -bounded and M' -Lipschitz over \mathbf{C} in probability via Lemma B.5. Then it follows that

$$\hat{g}_n^{\mathbf{S}'} \leq \hat{g}_n^{\mathbf{C}} \leq B' + M' \sup_{\mathbf{y}_1, \mathbf{y}_2 \in \mathbf{S}} \|\mathbf{y}_1 - \mathbf{y}_2\| =: B$$

in probability as $n \rightarrow \infty$.

- The function $\{(g_0 - g_0^*)(\hat{g}_n - g_0^*)\}(\cdot)$ is bounded and Lipschitz over \mathbf{S}' in probability

(where the constants do not depend on n). So is $\{(g_0 - g_0^*)(\hat{g}_n^{S'} - g_0^*)\}(\cdot)$ over \mathbf{S} . This also means that $\{(g_0 - g_0^*)(\hat{g}_n^{S'} - g_0^*)\}(\cdot)$ is equicontinuous over \mathbf{S} .

5. Returning to the quantity we mentioned at the end of Step 3, we note that

$$\begin{aligned}
& \frac{1}{m} \sum_{i=1}^m (g_0(\mathbf{x}_i) - g_0^*(\mathbf{x}_i))(\hat{a}_i - g_0^*(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \in \mathbf{S}'\}}) \\
&= \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\hat{g}_n^{S'} - g_0^*) \right)(\mathbf{x}_i) - \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\hat{g}_n^{S'} - g_0^*) \right)(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \notin \mathbf{S}'\}} \\
&= \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\hat{g}_n^{S'} - g_0^*) \right)(\mathbf{x}_i) - \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\hat{g}_n - g_0^*) \right)(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \notin \mathbf{S}'\}} \\
&\quad - \frac{1}{m} \sum_{i=1}^m \left((g_0 - g_0^*)(\hat{g}_n - \hat{g}_n^{S'}) \right)(\mathbf{x}_i)\mathbf{1}_{\{\mathbf{x}_i \notin \mathbf{S}'\}} \\
&= \text{(I)} + \text{(II)} + \text{(III)}.
\end{aligned}$$

We deal with each of these items separately.

- By the third fact listed in the above Step 4 and Theorem 3.1 of Rao (1962), (I) in the limit (i.e. as $n \rightarrow \infty$) is at most

$$\sup_{g \in G_2} \int_{\mathbf{S}} \{(g_0(\mathbf{x}) - g_0^*(\mathbf{x}))\{g(\mathbf{x}) - g_0^*(\mathbf{x})\}q(\mathbf{x})d\mathbf{x} \leq 0.$$

Note that g_0^* minimizes

$$\mathcal{G}(g) := \int_{\mathbf{S}} (g_0(\mathbf{x}) - g(\mathbf{x}))^2 q(\mathbf{x}) d\mathbf{x}$$

over all $g \in G_2$. The previous inequality thus follows by studying the functional derivative for the function $\mathcal{G}(\cdot)$ at g_0^* in the direction of $g - g_0^*$ (N.B. $g_0^* + \epsilon(g - g_0^*) \in G_2$ for $\epsilon \rightarrow 0$) for all $g \in G_2$.

- Both $|\text{(II)}|$ and $|\text{(III)}|$ in the limit can be arbitrarily small for \mathbf{S}' sufficiently close to \mathbf{S} . This follows from Cauchy–Schwarz inequality and an argument similar to that in Step 1.

6. We now put things together by noting that in light of Steps 1 to 5, for any ϵ , we can find

some \mathbf{S}' such that the quantity in (B.3) is no bigger than ϵ in probability as $n \rightarrow \infty$. Since the quantity in (B.3) is also non-negative, our claim that (B.3) converges to zero in probability is verified.

Finally, uniform consistency over any \mathbf{C} can be shown using exactly the same approach we demonstrated in the final stage of proving the first part of Theorem 2 via Lemma B.4. \square

B.4.2 Proof of Theorem 5

Proof. Our proof can be divided into three parts.

1. The case of $g_0 = 0$.

Using the definition of SCKLS in Appendix B.2, it is easy to verify that $T_n = \|\hat{g}_n - \tilde{g}_n\|_{n,m}$. For reasons that will become clear later, we denote \hat{g}_n° and \tilde{g}_n° the SCKLS and LL estimators based on the same covariates, evaluation points and bandwidth used in calculating T_n , but with the response vector $(\epsilon_1, \dots, \epsilon_n)'$ (instead of \mathbf{y}_n) and set $T_n^\circ = \|\tilde{g}_n^\circ - \hat{g}_n^\circ\|_{n,m}$. Obviously, when $g_0 = 0$ (which is the case here), $\hat{g}_n^\circ = \hat{g}_n, \tilde{g}_n^\circ = \tilde{g}_n$ and $T_n^\circ = T_n$.

Now, for $k = 1, \dots, B$, $T_{nk} = \|\hat{g}_{nk} - \tilde{g}_{nk}\|_{n,m}$, where \hat{g}_{nk} and \tilde{g}_{nk} are respectively the SCKLS and LL estimators based on the same covariates, evaluation points and bandwidth used in calculating T_n , but with the response vector $(u_{1k}\tilde{\epsilon}_1, \dots, u_{nk}\tilde{\epsilon}_n)'$. Further, we define a slightly modified bootstrap version of the test statistic as $T_{nk}^\circ = \|\hat{g}_{nk}^\circ - \tilde{g}_{nk}^\circ\|_{n,m}$, where \hat{g}_{nk}° and \tilde{g}_{nk}° are the SCKLS and LL estimators based on the same covariates, evaluation points and bandwidth used in calculating T_n , but with the response $(u_{1k}\epsilon_1, \dots, u_{nk}\epsilon_n)'$. Let $\mathbf{e} = (|\epsilon_1|, \dots, |\epsilon_n|)'$ and denote $p_n^\circ = \frac{1}{B} \sum_{i=1}^B \mathbf{1}_{\{T_n^\circ \leq T_{ni}^\circ\}}$. Then, it follows from the symmetry of the error distribution that conditioning on the values of the absolute errors (i.e. $(|\epsilon_1|, \dots, |\epsilon_n|)' = \mathbf{e}$), the quantities

$$T_n^\circ, T_{n1}^\circ, \dots, T_{nB}^\circ$$

are exchangeable. Consequently, as $B \rightarrow \infty$,

$$P(p_n^\circ \leq \alpha) = E\left\{P\left(p_n^\circ \leq \alpha \mid (|\epsilon_1|, \dots, |\epsilon_n|)' = \mathbf{e}\right)\right\} \leq \frac{\lfloor B\alpha \rfloor + 1}{1 + B} \rightarrow \alpha.$$

Back to the elements in the quantity p_n , our aim is to show that $\mathbf{1}_{\{T_n \leq T_{nk}^\circ\}} \leq \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}}$ for large n . Note that

$$T_{nk} - T_{nk}^\circ = \|\tilde{g}_{nk} - \hat{g}_{nk}\|_{n,m} - \|\tilde{g}_{nk}^\circ - \hat{g}_{nk}^\circ\|_{n,m} \leq \|\tilde{g}_{nk} - \hat{g}_{nk}^\circ\|_{n,m} - \|\tilde{g}_{nk}^\circ - \hat{g}_{nk}^\circ\|_{n,m} \leq \|\tilde{g}_{nk} - \tilde{g}_{nk}^\circ\|_{n,m}$$

Because we estimated the error vector in Step 1 using LL (without any shape restrictions), it follows from Proposition 7 of Fan and Guerre (2016) that $\sup_j |\tilde{\epsilon}_j - \epsilon_j| \leq O_p(n^{-2/(4+d)} \log^{1/2} n)$. By the linearity of the LL estimator (w.r.t. the response vector), we have that $\sup_k \|\tilde{g}_{nk} - \tilde{g}_{nk}^\circ\|_{n,m}^2 = O_p(n^{-4/(4+d)} \log n)$. Consequently, with arbitrarily high probability,

$$\inf_{k=1,\dots,B} (T_{nk} + \Delta_n - T_{nk}^\circ) > 0$$

for sufficiently large n . This yields $\mathbf{1}_{\{T_n^\circ \leq T_{nk}^\circ\}} \leq \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}}$ and thus $p_n \geq p_n^\circ$. As a result, $P(p_n \leq \alpha) \leq P(p_n^\circ \leq \alpha) \leq \alpha$, as required.

2. The general case of $g_0 \in G_2$.

To relate T_n to what we investigated before (i.e. $g_0 = 0$), we recall the definitions of \hat{g}_n° and \tilde{g}_n° from the previous case, and define an additional quantity \tilde{g}_n^\dagger to be the LL estimator in exactly the same setting, but is obtained using the response vector $(g_0(\mathbf{X}_1), \dots, g_0(\mathbf{X}_n))'$. By the linearity of the LL, $\tilde{g}_n = \tilde{g}_n^\circ + \tilde{g}_n^\dagger$. Since g_0 is continuously twice-differentiable, we have that

$$T_n = \|\tilde{g}_n - \hat{g}_n\|_{n,m} \leq \|\tilde{g}_n^\circ + \tilde{g}_n^\dagger - \hat{g}_n^\circ - g_0\|_{n,m} \leq \|\tilde{g}_n^\circ - \hat{g}_n^\circ\|_{n,m} + \|\tilde{g}_n^\dagger - g_0\|_{n,m} = T_n^\circ + O_p(h^2).$$

As a result, with arbitrarily high probability, for every $k = 1, \dots, B$,

$$T_{nk} + \Delta_n - T_n = T_{nk}^\circ - T_n^\circ + (T_{nk} - T_{nk}^\circ) - (T_n - T_n^\circ) + \Delta_n \geq T_{nk}^\circ - T_n^\circ$$

for sufficiently large n . This also leads to $\mathbf{1}_{\{T_n^\circ \leq T_{nk}^\circ\}} \leq \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}}$. We could then directly apply the argument from the previous case to conclude that $P(p_n \leq \alpha) \leq \alpha$.

3. The case of $g_0 \notin G_2$

Here g_0 is assumed to be fixed and continuously twice-differentiable.

First, two situations are considered.

- Under Assumption 2(i), we recall that

$$g_0^* := \operatorname{argmin}_{g \in G_2} \int_{\mathcal{S}} \{g(\mathbf{x}) - g_0(\mathbf{x})\}^2 Q(d\mathbf{x}).$$

Since $g_0 \notin G_2$, there must exist some compact set $\mathcal{S}' \subset \operatorname{int}(\mathcal{S})$ such that $Q(\mathcal{S}') > 0$ and

$$\inf_{\mathbf{x} \in \mathcal{S}'} |g_0^*(\mathbf{x}) - g_0(\mathbf{x})| > \delta.$$

Note that

$$T_n^2 = \|\hat{g}_n - \tilde{g}_n\|_{n,m}^2 \geq \frac{1}{m} \sum_{i=1}^m \left(\hat{g}_n(\mathbf{x}_i) - \tilde{g}_n(\mathbf{x}_i), \left(\frac{\partial(g_1 - g_2)}{\partial \mathbf{x}}(\mathbf{x}_i) \right)' h \right) \Sigma_i \begin{pmatrix} \hat{g}_n(\mathbf{x}_i) - \tilde{g}_n(\mathbf{x}_i) \\ \frac{\partial(\hat{g}_n - \tilde{g}_n)}{\partial \mathbf{x}}(\mathbf{x}_i) h \end{pmatrix} \mathbf{1}_{\{\mathbf{x}_i \in \mathcal{S}'\}}.$$

Here we have that $\tilde{g}_n \rightarrow g_0$ by Fan and Guerre (2016) and $\hat{g}_n \rightarrow g_0^*$ over \mathcal{S}' by our Theorem 4. Since $\tilde{g}_n - \hat{g}_n$ is Lipschitz over \mathcal{S}' , it is easy to verify (see also Step 3 of the proof of Theorem 4) that the righthand side of the above display equation is bounded below by $\delta^2 Q(\mathcal{S}')$ in the limit as $n \rightarrow \infty$ (also $h \rightarrow 0$). Consequently, $T_n \geq c'$ in probability for some $c' > 0$.

- Now under Assumption 2(ii), since $g_0 \notin G_2$ and the evaluation points are reasonably well spread across \mathcal{S} (i.e. Assumption 2(ii)), for sufficiently large and fixed m , we can always find some evaluation points where the imposed shape constraint is violated. This means that

$$\inf_{g \in G_2} \|g - g_0\|_{n,m} \geq c$$

in probability for some $c > 0$. So we still have that

$$T_n = \|\hat{g}_n - \tilde{g}_n\|_{n,m} \geq \|\hat{g}_n - g_0\|_{n,m} - \|\tilde{g}_n - g_0\|_{n,m} \geq \inf_{g \in G_2} \|g - g_0\|_{n,m} - o_p(1) \geq c'$$

in probability for some $c' > 0$.

Second, it follows from the proof for the case of $g_0 = 0$ that

$$T_{nk} = T_{nk}^\circ + T_{nk} - T_{nk}^\circ \leq \|\tilde{g}_{nk}^\circ\|_{n,m} + \|\tilde{g}_{nk} - \tilde{g}_{nk}^\circ\|_{n,m} = o_p(1).$$

Finally, write $W_{nk} = \mathbf{1}_{\{T_{nk} + \Delta_n > c'/2\}}$. We note that W_{n1}, \dots, W_{nB} are exchangeable. Thus, for any $\alpha \in (0, 1)$, as $n \rightarrow \infty$,

$$\begin{aligned} P(\text{Do not reject } H_0) &= P\left(\frac{1}{B} \sum_{k=1}^B \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}} \geq \alpha\right) \\ &\leq P(T_n \leq c'/2) + P\left(T_n > c'/2, \frac{1}{B} \sum_{k=1}^B \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}} \geq \alpha\right) \\ &\leq P(T_n \leq c'/2) + P\left(\frac{1}{B} \sum_{k=1}^B W_{nk} \geq \alpha\right) \\ &\leq P(T_n \leq c'/2) + \frac{E(W_{n1})}{\alpha} \rightarrow 0, \end{aligned}$$

where we used Markov's inequality in the final line above. So the Type II error at the alternative indeed converges to 0. □

B.5 Proof of Propositions in Appendix A.3

B.5.1 Proof of Proposition A.1

Proof. In view of Assumption 1 (v), for any sufficiently small \mathbf{h} , we have

$$K\left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}}\right) = \begin{cases} 0 & \text{if } \mathbf{x}_i \neq \mathbf{X}_j, \\ K(\mathbf{0}) & \text{if } \mathbf{x}_i = \mathbf{X}_j, \end{cases} \text{ for } \forall i, j.$$

Then, the objective function of (3) is equal to $\sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0})$, and thus

$$\operatorname{argmin}_{a_1, b_1, \dots, a_n, b_n} \sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0}) = \operatorname{argmin}_{a_1, \dots, a_n} \sum_{j=1}^n (y_j - a_j)^2$$

Writing $a_j = \alpha_j + \beta_j' \mathbf{X}_j$ and $\mathbf{b}_j = \beta_j$ for $j = 1, \dots, n$ by definition. Then, quadratic programming problem (3) can be rewritten as follows:

$$\begin{aligned} \min_{\alpha, \beta} \quad & \sum_{j=1}^n (y_j - (\alpha_j + \beta_j' \mathbf{X}_j))^2 \\ \text{subject to} \quad & \alpha_j + \beta_j' \mathbf{X}_j \leq \alpha_l + \beta_l' \mathbf{X}_j, \quad j, l = 1, \dots, n \\ & \beta_j \geq 0, \quad j = 1, \dots, n \end{aligned}$$

which is equivalent to the formulation of the CNLS estimator (A.5). \square

B.5.2 Proof of Proposition A.2

Proof. When $\min_{k=1, \dots, d} h_k \rightarrow \infty$, we have

$$K\left(\frac{\mathbf{X}_j - \mathbf{x}_i}{h}\right) = K(\mathbf{0}) \quad \text{for } \forall i, j. \quad (\text{B.8})$$

By substituting (B.8) into the objective function of (3) converges to

$$\sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 K(\mathbf{0}).$$

Next, we derive the minimum of the objective function in the limit. Let's consider

$$\operatorname{argmin}_{a_1, \mathbf{b}_1, \dots, a_m, \mathbf{b}_m} \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 \quad (\text{B.9})$$

subject to constraints. Rewrite $a_i + (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i = \alpha_i + \beta_i' \mathbf{X}_j$ for $i = 1, \dots, m$ and $j = 1, \dots, n$. Then the objective function of (3) can be rewritten as follows with (B.9).

$$\begin{aligned} \min_{\alpha_1, \beta_1, \dots, \alpha_m, \beta_m} \quad & \sum_{i=1}^m \sum_{j=1}^n (y_j - (\alpha_i + \beta_i' \mathbf{X}_j))^2 \\ \text{subject to} \quad & \alpha_i + \beta_i' \mathbf{x}_i \leq \alpha_l + \beta_l' \mathbf{x}_i \quad i, l = 1, \dots, m \\ & \beta_i \geq 0 \quad i = 1, \dots, m \end{aligned}$$

Here, since we do not impose any weight on the objective function, it is easy to see that

$\alpha_1 = \dots = \alpha_m$ and $\beta_1 = \dots = \beta_m$. Then the Afriat constraints become redundant, resulting in

$$\begin{aligned} \min_{\alpha, \beta} \quad & \sum_{j=1}^n (y_j - (\alpha + \beta' \mathbf{X}_j))^2 \\ \text{subject to} \quad & \beta \geq 0. \end{aligned}$$

□

B.5.3 Proof of Proposition A.3

Proof. In view of Assumption 1 (v), for any sufficiently small \mathbf{h} , we have

$$K\left(\frac{\mathbf{X}_j - \mathbf{x}_i}{\mathbf{h}}\right) = \begin{cases} 0 & \text{if } \mathbf{x}_i \neq \mathbf{X}_j, \\ K(\mathbf{0}) & \text{if } \mathbf{x}_i = \mathbf{X}_j, \end{cases} \text{ for } \forall i, j.$$

Then, the objective function of the SCKLS estimator (3) is equal to $\sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0})$, and thus

$$\operatorname{argmin}_{a_1, \mathbf{b}_1, \dots, a_n, \mathbf{b}_n} \sum_{j=1}^n (y_j - a_j)^2 K(\mathbf{0}) = \operatorname{argmin}_{a_1, \dots, a_n} \sum_{j=1}^n (y_j - a_j)^2$$

Also consider Assumption A1 (i) from Du et al. (2013), we can say something similar for CWB in y -space. For any sufficiently small \mathbf{h} , we have

$$A_j(\mathbf{x}_i) = \begin{cases} 0 & \text{if } \mathbf{x}_i \neq \mathbf{X}_j, \\ n & \text{if } \mathbf{x}_i = \mathbf{X}_j, \end{cases} \text{ for } \forall i, j.$$

and thus

$$\hat{g}(\mathbf{x}_i | \mathbf{p}) = \sum_{j=1}^n p_j A_j(\mathbf{X}_i) y_j = n p_i y_i \quad \forall i = 1, \dots, n. \quad (\text{B.10})$$

Then we can rewrite the CWB in y -space estimator as follows:

$$\begin{aligned} \min_{\mathbf{p}} \quad & D_y(\mathbf{p}) = \sum_{i=1}^n (y_i - n p_i y_i)^2 \\ \text{subject to} \quad & l(\mathbf{x}_i) \leq \hat{g}^{(s)}(\mathbf{x}_i | \mathbf{p}) \leq u(\mathbf{x}_i), \quad i = 1, \dots, n. \end{aligned} \quad (\text{B.11})$$

Recognize that if $\hat{g}_n = np_i y_i$ is true, then SCKLS and CWB in y-space are equivalent. Take \hat{g}_n as the solution to SCKLS estimator and let p_i be a set of decision variables, we see $\hat{g}_n = np_i y_i$ is simply a system of n equations and n unknowns. \square

C Testing for affinity using SCKLS

C.1 The procedure

To further illustrate the usefulness of SCKLS for testing other shapes, we study the problem of testing

$$H_0 : g_0 : \mathcal{S} \rightarrow \mathbb{R} \text{ is affine} \quad \text{against} \quad H_1 : g_0 : \mathcal{S} \rightarrow \mathbb{R} \text{ is not affine.}$$

The main idea of our test is motivated by Sen and Meyer (2017). The critical value of the test can be easily computed using Monte Carlo or bootstrap methods.

To start of with, we define \hat{g}_n^V , the SCKLS estimator with only a set of convexity constraints as

$$\begin{aligned} \min_{a_i, \mathbf{b}_i} \quad & \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 K \left(\frac{\|\mathbf{X}_j - \mathbf{x}_i\|}{h} \right) \\ \text{subject to} \quad & a_i - a_l \leq \mathbf{b}_i'(\mathbf{x}_i - \mathbf{x}_l), \quad i, l = 1, \dots, m \end{aligned}$$

Furthermore, \hat{g}_n^A , the SCKLS estimator using only a set of concavity constraints is defined as

$$\begin{aligned} \min_{a_i, \mathbf{b}_i} \quad & \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\mathbf{X}_j - \mathbf{x}_i)' \mathbf{b}_i)^2 K \left(\frac{\|\mathbf{X}_j - \mathbf{x}_i\|}{h} \right) \\ \text{subject to} \quad & a_i - a_l \geq \mathbf{b}_i'(\mathbf{x}_i - \mathbf{x}_l), \quad i, l = 1, \dots, m \end{aligned}$$

We now describe our testing procedure as follows.

1. First, we run linear regression on the response against the covariates and call the least squares fit g_n^L . Next, we fit the data using SCKLS (with evaluation points at $\mathbf{x}_1, \dots, \mathbf{x}_m$ and bandwidth h_n). The resulting estimators are denoted by \hat{g}_n^V and \hat{g}_n^A , where \hat{g}_n^V is the SCKLS estimator using only a set of convexity constraints, while

\hat{g}_n^Λ is the SCKLS estimator using only a set of concavity constraints, all based on $\{\mathbf{X}_j, y_j\}_{j=1}^n$. We then define the test statistics to be

$$T_n = \max \left[\frac{1}{m} \sum_{i=1}^m \{\hat{g}_n^V(\mathbf{x}_i) - g_n^L(\mathbf{x}_i)\}^2, \frac{1}{m} \sum_{i=1}^m \{\hat{g}_n^\Lambda(\mathbf{x}_i) - g_n^L(\mathbf{x}_i)\}^2 \right].$$

2. We simulate the distributional behavior of the test statistics B times under H_0 . For $k = 1, \dots, B$, we set the observations to be $\{\mathbf{X}_j, y_{jk}\}_{j=1}^n$ (i.e. no change in the values of the covariates), where $\mathbf{y}_{nk} = (y_{1k}, \dots, y_{nk})'$ is drawn using the wild bootstrap procedure as described in Section 4.2 (or the ordinary bootstrap procedure if we know that the errors are homogeneous). Then we run linear regression on \mathbf{y}_{nk} against the covariates and denote the least squares fit by g_{nk}^L . Fitting the data using SCKLS (with the same set of evaluation points and the same bandwidth as before) leads to the resulting estimators \hat{g}_{nk}^V and \hat{g}_{nk}^Λ , where \hat{g}_{nk}^V is the SCKLS estimator using only the convexity constraint, while \hat{g}_{nk}^Λ is the SCKLS estimator using only the concavity constraint, all based on $\{\mathbf{X}_j, y_{jk}\}_{j=1}^n$. So

$$T_{nk} = \max \left[\frac{1}{m} \sum_{i=1}^m \{\hat{g}_{nk}^V(\mathbf{x}_i) - g_{nk}^L(\mathbf{x}_i)\}^2, \frac{1}{m} \sum_{i=1}^m \{\hat{g}_{nk}^\Lambda(\mathbf{x}_i) - g_{nk}^L(\mathbf{x}_i)\}^2 \right].$$

3. The Monte Carlo p -value is defined as

$$p_n = \frac{1}{B} \sum_{k=1}^B \mathbf{1}_{\{T_n \leq T_{nk}\}}.$$

For a test of size $\alpha \in (0, 1)$, we reject H_0 if $p_n < \alpha$.

The intuition of the test is as follows. First, an affine function is both convex and concave. Therefore under H_0 , both SCKLS estimates, \hat{g}_n^V and \hat{g}_n^Λ , should be close to the linear fit g_n^L , so the value of T_n should be small. Second, a function is both convex and concave only if it is affine. So given enough observations, we should be able to reject the null hypothesis under H_1 . Third, we used the fact that T_n based on $\{\mathbf{X}_j, y_j\}_{j=1}^n$ and $\{\mathbf{X}_j, \epsilon_j\}_{j=1}^n$ are exactly the same under H_0 when simulating the distributional behavior of T_n .

Finally, we remark that in case we know that g_0 is monotonically increasing a priori, we could test $H'_0 : g_0$ is monotonically increasing and affine using essentially the same procedure with only minor modifications described in the following: we instead run linear regression with signed constraints in both Step 1 and Step 2, replace \hat{g}_n^V by the SCKLS with both the convexity and monotonicity constraints, and replace \hat{g}_n^A by the SCKLS with both the concavity and monotonicity constraints.

C.2 A simulation study

We now examine the finite-sample performance of the affinity test using data generated from the following DGP:

$$g_0(\mathbf{x}) = \frac{1}{d} \sum_{k=1}^d x_k^p \quad (\text{C.1})$$

where $\mathbf{x} = (x_1, \dots, x_d)'$. With n observations, for each pair (\mathbf{X}_j, y_j) , each component of the input, \mathbf{X}_{jk} , is randomly and independently drawn from uniform distribution $\text{unif}[0, 1]$, and the additive noise, ϵ_j , is randomly and independently sampled from a normal distribution, $N(0, 0.1)$.

We considered different sample sizes $n \in \{100, 300, 500\}$ and vary the number of inputs $d \in \{1, 2\}$, and perform 100 simulations to compute the rejection rate for each scenario. We used the ordinary bootstrap method with $B = 500$.

In the scenarios we considered g_0 is affine if $p = 1.0$, and is non-linear if $p \in \{0.2, 0.5, 2, 5\}$. Table C.1 show the rejection rate for each scenario with one-input and two-input at $\alpha = 0.05$. We conclude that the proposed test works well with a moderate sample size.

D An algorithm for SCKLS computational performance

For a given number of evaluation points, m , SCKLS requires $m(m-1)$ concavity constraints. Larger values of m provide a more flexible functional estimate, but also increase the number of constraints quadratically, thus, the amount of time needed to solve the quadratic program also increases quadratically. Since one can select the number of evaluation points in SCKLS, by selecting m the computational complexity can be potentially reduced relative to CNLS

Table C.1. Rejection rate of the affinity test using SCKLS at $\alpha = 0.05$

Sample size (n)	Shape Parameter (p)	Power of the Test	
		$d = 1$	$d = 2$
100	0.2	0.99	0.74
	0.5	0.97	0.79
	1.0	0.05	0.02
	2.0	1.00	1.00
	5.0	1.00	1.00
300	0.2	1.00	1.00
	0.5	1.00	0.99
	1.0	0.05	0.01
	2.0	1.00	1.00
	5.0	1.00	1.00
500	0.2	1.00	1.00
	0.5	1.00	1.00
	1.0	0.08	0.01
	2.0	1.00	1.00
	5.0	1.00	1.00

or estimates on denser grids, i.e. with $m(m - 1) \ll n(n - 1)$.

Further, Dantzig et al. (1954, 1959) proposed an iterative approach that reduces the size of large-scale problems by relaxing a subset of the constraints and solving the relaxed model with only a subset V of constraints, checking which of the excluded constraints are violated, and iteratively adding violated constraints to the relaxed model until an optimal solution satisfies all constraints. Lee et al. (2013), who applied the approach to CNLS, found a significant reduction in computational time. Computational performances also improves if a subset of the constraints can be identified which are likely to be needed in the model. Lee et al. (2013) find the concavity constraints corresponding to pairs of observations that are close in terms of the ℓ_2 norm measured over input vectors and more likely to be binding than those corresponding to the distant observations. We use this insight to develop a strategy for identifying constraints to include in the initial subset V , when solving SCKLS as described below.

Given a grid to evaluate the constraints of the SCKLS estimator, we define the initial subset of constraints V as those constraints constructed by adjacent grid points as shown

in Figure D.1. Further, we summarize our implementation of the algorithm proposed in Lee et al. (2013) below and label it as Algorithm 1.

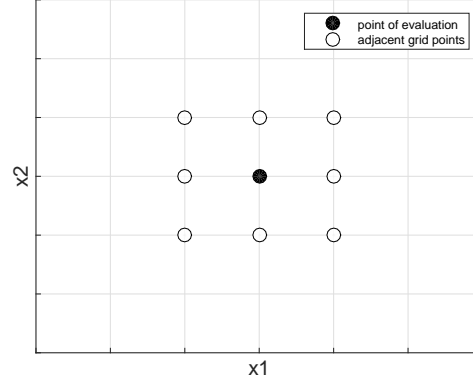


Figure D.1. Definition of adjacent grid in two-dimensional case.

Algorithm 1 Iterative approach for SCKLS computational speedup

```

 $t \leftarrow 0$ 
 $V \leftarrow \{(i, l) : \mathbf{x}_i \text{ and } \mathbf{x}_l \text{ are adjacent, } i < l\}$ 
Solve relaxed SCKLS with  $V$  to find initial solution  $\{a_i^{(0)}, \mathbf{b}_i^{(0)}\}_{i=1}^m$ 
while  $\{a_i^{(t)}, \mathbf{b}_i^{(t)}\}_{i=1}^m$  satisfies all constraints in (3) do
     $t \leftarrow t + 1$ 
     $U \leftarrow \{(i, l) : \mathbf{x}_i \text{ and } \mathbf{x}_l \text{ do not satisfy constraints in (3)}\}$ 
     $V \leftarrow V \cup U$ 
    Solve relaxed SCKLS with  $V$  to find solution  $\{a_i^{(t)}, \mathbf{b}_i^{(t)}\}_{i=1}^m$ 
end while
return  $\{a_i^{(t)}, \mathbf{b}_i^{(t)}\}_{i=1}^m$ 

```

E Comprehensive results of existing and additional numerical experiments

We show the comprehensive results of experiments in Section 5 and additional experiments to show the performance of the SCKLS estimator and its extensions. For the CWB estimator, we use the convex optimization solver **SeDuMi** because **quadprog** was not able to solve CWB⁶.

For CWB estimator, we use a local linear estimator to obtain the weighting matrix $A_j(\mathbf{x})$ in (A.6). The first partial derivative of $\hat{g}(\mathbf{x}|\mathbf{p})$ is obtained by approximating the derivatives through numerical differentiation $\hat{g}^{(1)}(\mathbf{x}|\mathbf{p}) = \frac{\hat{g}(\mathbf{x}+\Delta|\mathbf{p}) - \hat{g}(\mathbf{x}|\mathbf{p})}{\Delta}$, where Δ is a small positive constant⁷.

E.1 Uniform input – high signal-to-noise ratio (Experiment 1)

We compare the following seven estimators: SCKLS with fixed bandwidth, SCKLS with variable bandwidth, CNLS, CWB in p -space and CWB in y -space, LL, and parametric Cobb–Douglas function estimated via ordinary least squares (OLS). Table E.1 and Table E.2 show the RMSE of Experiment 1 on observation points and evaluation points respectively.

Table E.3 shows the computational time of Experiment 1 for each estimator.

We also conduct simulations with different bandwidths to analyze the sensitivity of each estimator to bandwidths. We estimate SCKLS with fixed bandwidth, CWB in p -space and local linear with bandwidth $h \in [0, 10]$ with an increment by 0.01 for 1-input setting, and we use bandwidth $\mathbf{h} \in [0, 5] \times [0, 5]$ with an increment by 0.25 for 2-input setting. We perform 100 simulations for each bandwidth, and compute the optimal bandwidth with LOOCV for each simulation. Figure 1 displays the average RMSE of each estimator. The distribution of bandwidths selected by LOOCV are shown in the histogram. The instances

⁶For CWB, **SeDuMi** provides a better solution than **quadprog**, while both **SeDuMi** and **quadprog** give exactly the same solution for SCKLS.

⁷Du et al. (2013) proposes to use an analytical derivative for the first partial derivative of $\hat{g}(\mathbf{x}|\mathbf{p})$; however, the analytical derivative performs similarly to numerical differentiation as shown in Racine (2016). We propose two alternative methods to compute the first partial derivative, and compared them in Appendix A.2.2.

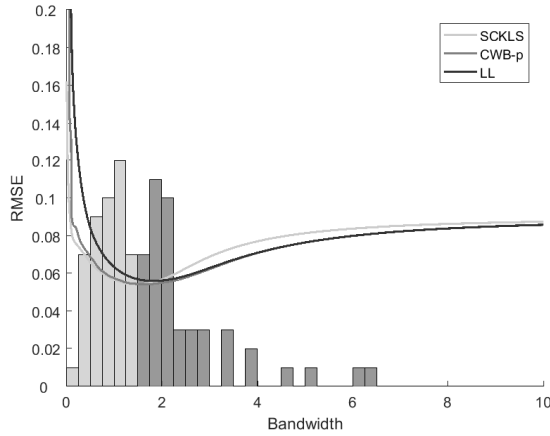
Table E.1. RMSE on observation points for Experiment 1

Number of observations		Average of RMSE on observation points				
		100	200	300	400	500
2-input	SCKLS fixed bandwidth	0.193	0.171	0.141	0.132	0.118
	SCKLS variable bandwidth	0.183	0.158	0.116	0.118	0.098
	CNLS	0.229	0.163	0.137	0.138	0.116
	CWB in p -space	0.189	0.167	0.158	0.140	0.129
	CWB in y -space	0.205	0.136	0.173	0.141	0.120
	LL	0.212	0.166	0.149	0.152	0.140
	Cobb–Douglas	0.078	0.075	0.048	0.039	0.043
3-input	SCKLS fixed bandwidth	0.230	0.187	0.183	0.152	0.165
	SCKLS variable bandwidth	0.216	0.183	0.175	0.143	0.142
	CNLS	0.294	0.202	0.189	0.173	0.168
	CWB in p -space	0.228	0.221	0.210	0.183	0.172
	CWB in y -space	0.209	0.362	0.218	0.154	0.160
	LL	0.250	0.230	0.235	0.203	0.181
	Cobb–Douglas	0.104	0.089	0.070	0.047	0.041
4-input	SCKLS fixed bandwidth	0.225	0.248	0.228	0.203	0.198
	SCKLS variable bandwidth	0.217	0.219	0.210	0.180	0.179
	CNLS	0.315	0.294	0.246	0.235	0.214
	CWB in p -space	0.238	0.262	0.231	0.234	0.198
	CWB in y -space	0.222	0.240	0.248	0.303	0.332
	LL	0.256	0.297	0.252	0.240	0.226
	Cobb–Douglas	0.120	0.073	0.091	0.067	0.063

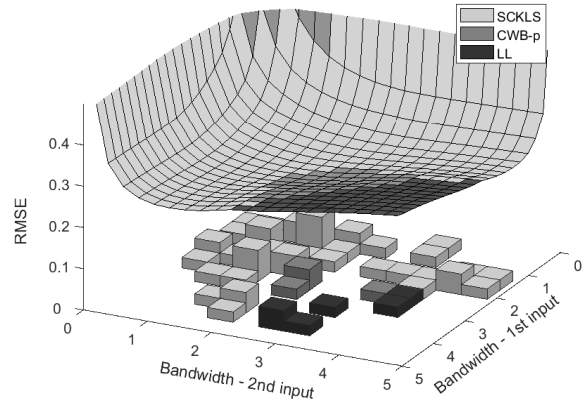
when SCKLS, CWB- p , and local linear provide the lowest RMSE are shown in light gray, gray and dark gray respectively on the histogram. For one-input scenario, the SCKLS and CWB estimator perform similar for bandwidth between 0.25 - 2.25 as shown by the closeness of the light gray and gray curves in (a). In contrast, for two-input scenario, the SCKLS estimator performs better for most of the LOOCV values as shown by the majority of the histogram colored in light gray. This indicates that LOOCV calculate for unconstrained estimator provide bandwidths that work well for the SCKLS estimator.

Table E.2. RMSE on evaluation points for Experiment 1

Number of observations		Average of RMSE on evaluation points				
		100	200	300	400	500
2-input	SCKLS fixed bandwidth	0.219	0.189	0.150	0.147	0.128
	SCKLS variable bandwidth	0.212	0.176	0.125	0.132	0.103
	CNLS	0.350	0.299	0.260	0.284	0.265
	CWB in p -space	0.206	0.186	0.174	0.154	0.143
	CWB in y -space	0.259	0.228	0.228	0.172	0.167
	LL	0.247	0.182	0.167	0.171	0.156
Cobb–Douglas		0.076	0.076	0.049	0.040	0.043
3-input	SCKLS fixed bandwidth	0.283	0.231	0.238	0.213	0.215
	SCKLS variable bandwidth	0.292	0.237	0.235	0.196	0.187
	CNLS	0.529	0.587	0.540	0.589	0.598
	CWB in p -space	0.291	0.289	0.269	0.252	0.233
	CWB in y -space	0.314	0.474	0.265	0.346	0.261
	LL	0.336	0.340	0.360	0.326	0.264
Cobb–Douglas		0.116	0.098	0.080	0.052	0.046
4-input	SCKLS fixed bandwidth	0.321	0.357	0.329	0.308	0.290
	SCKLS variable bandwidth	0.378	0.348	0.363	0.320	0.301
	CNLS	0.845	0.873	0.901	0.827	0.792
	CWB in p -space	0.360	0.385	0.358	0.361	0.325
	CWB in y -space	0.355	0.470	0.338	0.410	0.602
	LL	0.482	0.527	0.483	0.495	0.445
Cobb–Douglas		0.146	0.091	0.115	0.081	0.080



(a) One-input



(b) Two-input

Figure E.1. The histogram shows the distribution of bandwidths selected by LOOCV. The curves show the relative performance of each estimator.

Table E.3. Computational time for Experiment 1

Number of observations		Average of computational time in seconds; (percentage of Afriat constraints included in the final optimization problem)				
		100	200	300	400	500
2-input	SCKLS fixed bandwidth	14.1 (6.14%)	13.3 (5.28%)	42.2 (8.86%)	34.7 (7.80%)	77.4 (8.31%)
	SCKLS variable bandwidth	16.4 (3.47%)	33.9 (3.44%)	27.6 (3.34%)	36.0 (3.22%)	50.6 (3.53%)
	CNLS	2.0 (100%)	6.1 (100%)	16.5 (100%)	26.5 (100%)	55.3 (100%)
	CWB in p -space	24.1 (2.39%)	33.2 (2.35%)	76.6 (2.35%)	82.3 (2.35%)	130 (2.35%)
	CWB in y -space	39.3 (2.35%)	92.7 (2.35%)	111 (2.35%)	190 (2.35%)	233 (2.36%)
3-input	SCKLS fixed bandwidth	26.9 (16.0%)	40.4 (16.6%)	45.5 (16.3%)	67.3 (16.4%)	136 (16.2%)
	SCKLS variable bandwidth	20.0 (15.7%)	42.0 (15.9%)	37.4 (15.8%)	47.1 (15.8%)	58.2 (15.9%)
	CNLS	3.8 (100%)	16.4 (100%)	37.0 (100%)	82.9 (100%)	161 (100%)
	CWB in p -space	47.6 (15.5%)	71.5 (15.5%)	100 (15.5%)	202 (15.5%)	255 (15.5%)
	CWB in y -space	120 (15.5%)	357 (15.5%)	443 (15.5%)	529 (15.5%)	424 (15.5%)
4-input	SCKLS fixed bandwidth	47.5 (40.1%)	71.6 (39.9%)	77.4 (39.9%)	166 (40.0%)	235 (39.8%)
	SCKLS variable bandwidth	26.8 (39.9%)	45.6 (40.0%)	46.8 (39.8%)	60.5 (39.9%)	74.8 (39.8%)
	CNLS	5.8 (100%)	22.4 (100%)	79.1 (100%)	139.8 (100%)	287.8 (100%)
	CWB in p -space	68.8 (39.8%)	136 (39.8%)	196 (39.8%)	327 (39.8%)	442 (39.8%)
	CWB in y -space	91.3 (39.8%)	175 (39.8%)	195 (39.8%)	535 (39.8%)	545 (39.8%)

E.2 Uniform input – low signal-to-noise ratio

We consider a Cobb–Douglas production function with d -inputs and one-output,

$$g_0(x_1, \dots, x_d) = \prod_{k=1}^d x_k^{\frac{0.8}{d}}.$$

For each pair (\mathbf{X}_j, y_j) , each component of the input, \mathbf{X}_{jk} , is randomly and independently drawn from uniform distribution $unif[1, 10]$, and the additive noise, ϵ_j , is randomly and independently sampled from a normal distribution, $N(0, 1.3^2)$. We consider 15 different scenarios with different numbers of observations (100, 200, 300, 400 and 500) and input dimension (2, 3 and 4). The number of evaluation points is fixed at 400, and set as a uniform grid. This experiment has a higher noise level in the data generation process relative to Experiment 1.

We compare following seven estimators: SCKLS with fixed bandwidth, SCKLS with variable bandwidth, CNLS, CWB in p -space, CWB in y -space, LL, and parametric Cobb–Douglas function estimated via ordinary least squares (OLS). Table E.4 and Table E.5 show the RMSE of this experiment on observation points and evaluation points respectively.

Table E.4. RMSE on observation points for Experiment: uniform input with low signal-to-noise ratio

Number of observations		Average of RMSE on observation points				
		100	200	300	400	500
2-input	SCKLS fixed bandwidth	0.239	0.203	0.203	0.155	0.140
	SCKLS variable bandwidth	0.240	0.185	0.168	0.139	0.119
	CNLS	0.279	0.231	0.194	0.168	0.151
	CWB in p -space	0.314	0.215	0.237	0.275	0.151
	CWB in y -space	0.241	0.229	0.173	0.178	0.206
	LL	0.287	0.244	0.230	0.214	0.161
	Cobb–Douglas	0.109	0.108	0.081	0.042	0.048
3-input	SCKLS fixed bandwidth	0.292	0.263	0.221	0.204	0.184
	SCKLS variable bandwidth	0.281	0.242	0.198	0.180	0.175
	CNLS	0.379	0.303	0.275	0.224	0.214
	CWB in p -space	0.318	0.306	0.308	0.244	0.214
	CWB in y -space	0.281	0.273	0.225	0.320	0.271
	LL	0.333	0.306	0.288	0.259	0.214
	Cobb–Douglas	0.176	0.118	0.101	0.084	0.072
4-input	SCKLS fixed bandwidth	0.317	0.291	0.249	0.241	0.254
	SCKLS variable bandwidth	0.290	0.254	0.236	0.222	0.215
	CNLS	0.491	0.356	0.311	0.293	0.313
	CWB in p -space	0.400	0.318	0.273	0.260	0.289
	CWB in y -space	0.312	0.338	0.262	0.365	0.453
	LL	0.335	0.342	0.257	0.274	0.283
	Cobb–Douglas	0.157	0.150	0.112	0.075	0.077

Table E.5. RMSE on evaluation points for Experiment: uniform input with low signal-to-noise ratio

Number of observations		Average of RMSE on evaluation points				
		100	200	300	400	500
2-input	SCKLS fixed bandwidth	0.253	0.225	0.222	0.172	0.160
	SCKLS variable bandwidth	0.255	0.205	0.179	0.149	0.135
	CNLS	0.319	0.355	0.334	0.255	0.267
	CWB in p -space	0.329	0.239	0.262	0.305	0.177
	CWB in y -space	0.263	0.241	0.198	0.228	0.180
	LL	0.330	0.272	0.257	0.239	0.194
	Cobb–Douglas	0.112	0.112	0.083	0.044	0.049
3-input	SCKLS fixed bandwidth	0.367	0.339	0.302	0.268	0.231
	SCKLS variable bandwidth	0.364	0.303	0.256	0.230	0.224
	CNLS	0.743	0.778	0.744	0.696	0.620
	CWB in p -space	0.398	0.392	0.434	0.336	0.274
	CWB in y -space	0.401	0.473	0.385	0.450	0.525
	LL	0.452	0.444	0.438	0.398	0.302
	Cobb–Douglas	0.202	0.130	0.110	0.093	0.079
4-input	SCKLS fixed bandwidth	0.405	0.460	0.349	0.350	0.347
	SCKLS variable bandwidth	0.419	0.434	0.375	0.354	0.315
	CNLS	1.019	0.950	0.985	1.043	1.106
	CWB in p -space	0.514	0.520	0.393	0.390	0.452
	CWB in y -space	0.514	0.513	0.425	0.501	0.708
	LL	0.524	0.626	0.451	0.491	0.550
	Cobb–Douglas	0.187	0.194	0.134	0.092	0.091

E.3 Different numbers of evaluation points (Experiment 2)

We compare following four estimators: SCKLS with fixed bandwidth, SCKLS with variable bandwidth, CWB in p -space and CWB in y -space. Table E.6 and Table E.7 show the RMSEs of Experiment 2 on observation points and evaluation points respectively. In addition, Table E.8 shows the computational time of Experiment 2 for each estimator.

Table E.6. RMSE on observation points for Experiment 2

Number of evaluation points		Average of RMSE on observation points		
		100	300	500
2-input	SCKLS fixed bandwidth	0.142	0.141	0.141
	SCKLS variable bandwidth	0.113	0.112	0.112
	CWB in p -space	0.149	0.151	0.156
	CWB in y -space	0.225	0.122	0.129
3-input	SCKLS fixed bandwidth	0.198	0.203	0.197
	SCKLS variable bandwidth	0.169	0.167	0.166
	CWB in p -space	0.218	0.234	0.231
	CWB in y -space	0.345	0.241	0.222
4-input	SCKLS fixed bandwidth	0.239	0.207	0.206
	SCKLS variable bandwidth	0.195	0.192	0.191
	CWB in p -space	0.219	0.227	0.296
	CWB in y -space	0.466	0.290	0.292

Table E.7. RMSE on evaluation points for Experiment 2

Number of evaluation points		Average of RMSE on evaluation points		
		100	300	500
2-input	SCKLS fixed bandwidth	0.181	0.164	0.158
	SCKLS variable bandwidth	0.140	0.128	0.124
	CWB in p -space	0.195	0.180	0.179
	CWB in y -space	0.262	0.162	0.169
3-input	SCKLS fixed bandwidth	0.304	0.267	0.257
	SCKLS variable bandwidth	0.242	0.213	0.205
	CWB in p -space	0.332	0.329	0.302
	CWB in y -space	0.792	0.582	0.559
4-input	SCKLS fixed bandwidth	0.383	0.296	0.270
	SCKLS variable bandwidth	0.386	0.304	0.265
	CWB in p -space	0.403	0.359	0.415
	CWB in y -space	1.040	0.352	0.381

Table E.8. Computational time for Experiment 2

Number of evaluation points		Average of computational time in seconds; (percentage of Afriat constraints included in the final optimization)		
		100	300	500
2-input	SCKLS fixed bandwidth	26.6 (11.7%)	28.3 (6.6%)	34 (5.4%)
	SCKLS variable bandwidth	21.3 (9.9%)	21.6 (4.4%)	24.9 (3.2%)
	CWB in p -space	41 (8.8%)	56.5 (3.2%)	74.2 (2.0%)
	CWB in y -space	52.8 (8.8%)	103 (3.2%)	146 (2.0%)
3-input	SCKLS fixed bandwidth	84.8 (29.1%)	112 (16.7%)	134 (13.3%)
	SCKLS variable bandwidth	21.1 (28.5%)	37.2 (15.8%)	59.1 (12.4%)
	CWB in p -space	121 (28.2%)	221 (15.5%)	310 (12.2%)
	CWB in y -space	181 (28.2%)	625 (15.5%)	948 (12.2%)
4-input	SCKLS fixed bandwidth	149 (62.3%)	170 (40.0%)	597 (27.7%)
	SCKLS variable bandwidth	24.6 (62.1%)	52.7 (39.9%)	468 (27.5%)
	CWB in p -space	175 (61.9%)	275 (39.8%)	729 (27.4%)
	CWB in y -space	189 (61.9%)	288 (39.8%)	579 (27.4%)

E.4 Non-uniform input

Experiment 4. We consider a Cobb–Douglas production function with d -inputs and one-output,

$$g_0(x_1, \dots, x_d) = \prod_{k=1}^d x_k^{\frac{0.8}{d}}.$$

For each pair (\mathbf{X}_j, y_j) , each component of the input, \mathbf{X}_{jk} , is randomly and independently drawn from a truncated exponential distribution with density function

$$f(x) = \frac{3}{e^{-3} - e^{-30}} e^{-3x} \mathbf{1}_{\{x \in [1, 10]\}},$$

and the additive noise, ϵ_j , is randomly sampled from a normal distribution, $N(0, 0.7^2)$. We consider 15 different scenarios with different numbers of observations (100, 200, 300, 400 and 500) and input dimension (2, 3 and 4). The number of evaluation point is fixed at 400. Note that this experiment only differs from Experiment 1 in that the distribution of inputs is skewed and thus non-uniform.

We compare following seven estimators: SCKLS with fixed bandwidth with uniform/non-uniform grid, SCKLS with variable bandwidth with uniform/non-uniform grid, CNLS, CWB in p -space with uniform/non-uniform grid. These extension of SCKLS were presented in detail in Appendix A.1. Table E.9 and Table E.10 show the RMSEs of Experiment 4 on observation points and evaluation points respectively. A uniform grid is used like in Experiment 1. As the dimension of input space and the number of observations increase, SCKLS with variable bandwidth performs better than the fixed bandwidth estimator. SCKLS with non-uniform grid performs better than SCKLS with uniform grid for almost all scenarios, largely due to the fact that the DGP has non-uniform input. Consequently, we conclude that variable bandwidth methods, such as k -NN approach, and non-uniform grid could be useful to handle skewed input data which is a common feature of census manufacturing data which is the type of data we considered in the application of the main manuscript.

Table E.9. RMSE on observation points for Experiment: non-uniform input

Number of observations		Average of RMSE on observation points				
		100	200	300	400	500
2-input	SCKLS fixed/uniform	0.179	0.151	0.144	0.121	0.108
	SCKLS fixed/non-uniform	0.185	0.153	0.159	0.123	0.107
	SCKLS variable/uniform	0.183	0.156	0.142	0.125	0.104
	SCKLS variable/non-uniform	0.176	0.144	0.132	0.114	0.093
	CNLS	0.193	0.160	0.140	0.130	0.117
	CWB p -space/uniform	0.256	0.162	0.180	0.139	0.125
	CWB p -space/non-uniform	0.243	0.160	0.174	0.135	0.125
3-input	SCKLS fixed/uniform	0.197	0.184	0.172	0.164	0.167
	SCKLS fixed/non-uniform	0.200	0.181	0.173	0.161	0.172
	SCKLS variable/uniform	0.212	0.187	0.170	0.175	0.170
	SCKLS variable/non-uniform	0.210	0.180	0.162	0.160	0.155
	CNLS	0.303	0.246	0.201	0.185	0.166
	CWB p -space/uniform	0.243	0.436	0.173	0.174	0.184
	CWB p -space/non-uniform	0.233	0.194	0.176	0.165	0.173
4-input	SCKLS fixed/uniform	0.219	0.211	0.196	0.209	0.187
	SCKLS fixed/non-uniform	0.210	0.206	0.181	0.197	0.180
	SCKLS variable/uniform	0.208	0.193	0.167	0.171	0.170
	SCKLS variable/non-uniform	0.206	0.193	0.164	0.169	0.168
	CNLS	0.347	0.292	0.250	0.228	0.218
	CWB p -space/uniform	0.219	0.205	0.205	0.184	0.218
	CWB p -space/non-uniform	0.221	0.205	0.182	0.170	0.170

Table E.10. RMSE on evaluation points for Experiment: non-uniform input

Number of observations		Average of RMSE on evaluation points				
		100	200	300	400	500
2-input	SCKLS fixed/uniform	0.262	0.220	0.244	0.157	0.196
	SCKLS fixed/non-uniform	0.212	0.174	0.195	0.138	0.131
	SCKLS variable/uniform	0.246	0.204	0.192	0.142	0.136
	SCKLS variable/non-uniform	0.193	0.160	0.145	0.120	0.100
	CNLS	0.435	0.402	0.404	0.379	0.381
	CWB p -space/uniform	0.422	0.287	0.376	0.246	0.264
	CWB p -space/non-uniform	0.283	0.186	0.215	0.159	0.162
3-input	SCKLS fixed/uniform	0.323	0.308	0.311	0.286	0.293
	SCKLS fixed/non-uniform	0.268	0.254	0.259	0.235	0.249
	SCKLS variable/uniform	0.335	0.303	0.281	0.262	0.254
	SCKLS variable/non-uniform	0.278	0.243	0.219	0.212	0.196
	CNLS	0.828	0.824	0.828	0.786	0.782
	CWB p -space/uniform	0.438	0.684	0.357	0.363	0.350
	CWB p -space/non-uniform	0.315	0.265	0.257	0.235	0.242
4-input	SCKLS fixed/uniform	0.406	0.398	0.397	0.404	0.400
	SCKLS fixed/non-uniform	0.339	0.343	0.333	0.371	0.331
	SCKLS variable/uniform	0.417	0.423	0.368	0.364	0.356
	SCKLS variable/non-uniform	0.359	0.359	0.313	0.302	0.280
	CNLS	1.129	1.107	1.220	1.196	1.223
	CWB p -space/uniform	0.421	0.442	0.435	0.418	0.487
	CWB p -space/non-uniform	0.354	0.344	0.308	0.286	0.280

E.5 Estimation with a misspecified shape

We use the DGP proposed by Olesen and Ruggiero (2014) that is consistent with the regular ultra passum law (Frisch, 1964), which appears to have an “S”-shape.

$$g_0(x_1, x_2) = F(h(x_1, x_2))$$

where the scaling function is: $F(w) = \frac{15}{1+e^{-5 \log(w)}}$, and the linear homogeneous core function is

$$h(x_1, x_2) = \left(\beta x_1^{\frac{\sigma-1}{\sigma}} + (1-\beta) x_2^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}$$

with $\beta = 0.45$ and $\sigma = 1.51$. For $j = 1, \dots, n$, input, $\mathbf{X}_j = (X_{j1}, X_{j2})'$, is generated in polar coordinates with angles η and modulus ω independently uniformly distributed on $[0.05, \pi/2 - 0.05]$ and $[0, 2.5]$, respectively. The additive noise, ϵ_j , is randomly sampled from $N(0, 0.7^2)$.

Note that this DGP is not concave. Here we run this experiment to assess the performance of each estimator in case of shape misspecification. Table E.11 and Table E.12 show the RMSEs of this experiment on observation points and evaluation points. Figure E.2 shows the estimation results with 1-input S-shape function from a typical run of SCKLS. The figure shows that the SCKLS estimator results in a linear estimates for areas where concavity is violated. Here the CWB estimator performs slightly worse when the function is misspecified. We speculate that the main reason for this is that the optimization problem becomes too complicated to solve since intuitively there are many binding constraints when the data is generated by the misspecified functional form, and thus, it becomes hard for the solver to find a feasible solution and an improving direction.

Table E.11. RMSE on observation points for Experiment: misspecified shape

Number of observations	Average of RMSE on observation points				
	100	200	300	400	500
SCKLS fixed bandwidth	1.424	1.435	1.405	1.392	1.421
CNLS	1.326	1.346	1.337	1.316	1.353
CWB in p -space	6.310	6.731	6.602	5.909	6.110

Table E.12. RMSE on evaluation points for Experiment: misspecified shape

Number of observations	Average of RMSE on evaluation points				
	100	200	300	400	500
SCKLS fixed bandwidth	1.337	1.162	1.149	1.140	1.123
CNLS	1.375	1.424	1.404	1.403	1.385
CWB in p -space	9.100	9.483	9.599	8.435	8.719

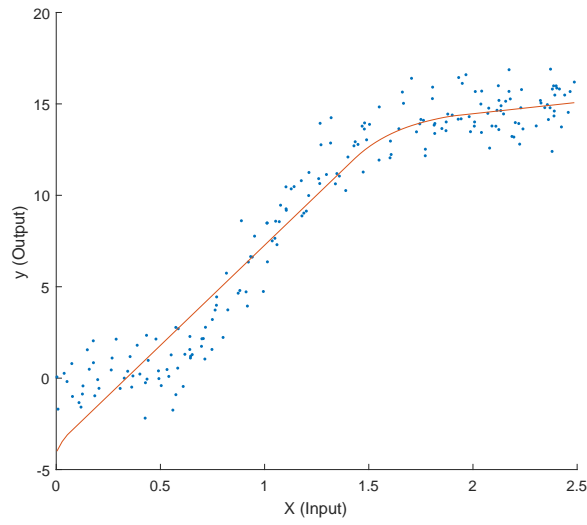


Figure E.2. A typical run of SCKLS when the truth is S-shaped.

F Semiparametric partially linear model

F.1 The procedure

We develop a semiparametric partially linear model including the SCKLS estimator and a linear function of contextual variables. The partially linear model is often used in practice. The model estimated is represented as follows:

$$y_j = \mathbf{Z}_j' \boldsymbol{\gamma} + g_0(\mathbf{X}_j) + \epsilon_j$$

where $\mathbf{Z}_j = (Z_{j1}, Z_{j2}, \dots, Z_{jl})'$ denotes contextual variables and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_l)'$ is the coefficient of contextual variables, see Johnson and Kuosmanen (2011, 2012). Then, we estimate the coefficient of contextual variable:

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^n \tilde{\mathbf{Z}}_j \tilde{\mathbf{Z}}_j' \right)^{-1} \left(\sum_{j=1}^n \tilde{\mathbf{Z}}_j \tilde{y}_j \right)$$

where $\tilde{\mathbf{Z}}_j = \mathbf{Z}_j - \hat{E}[\mathbf{Z}_j | \mathbf{X}_j]$ and $\tilde{y}_j = y_j - \hat{E}[y_j | \mathbf{X}_j]$ respectively, and each conditional expectation is estimated by kernel estimation method such as local linear. Finally, we apply the SCKLS estimator to the data $\{\mathbf{X}_j, y_j - \mathbf{Z}_j' \hat{\boldsymbol{\gamma}}\}_{j=1}^n$. Robinson (1988) proved that $\hat{\boldsymbol{\gamma}}$ is $n^{1/2}$ -consistent for $\boldsymbol{\gamma}$ and asymptotically normal under regularity conditions. For details of the partially linear model, see Li and Racine (2007).

F.2 A simulation study

We show the effect of adding contextual variables \mathbf{Z}_j to the estimation performance by comparing SCKLS with and without contextual variables. We use two different Cobb–Douglas production functions as the true DGP:

$$g_0(\mathbf{x}, z) = \prod_{k=1}^d x_k^{\frac{0.8}{d}} + z\gamma, \tag{F.1}$$

$$g_0(\mathbf{x}) = \prod_{k=1}^d x_k^{\frac{0.8}{d}}, \quad (\text{F.2})$$

where for each (\mathbf{X}_j, Z_j, y_j) , the contextual variable Z_j is a scalar value independent of \mathbf{X}_j drawn randomly and independently from $unif[0, 1]$, the coefficient of the contextual variable $\gamma = 5$, and other parameters follow DGP from Experiment 1. We apply SCKLS with and without contextual variables to the data generated by the true production function (F.1) and (F.2), respectively.

Table F.1 and Table F.2 show the RMSEs of this experiment on observation points and evaluation points respectively. The RMSE is obtained by comparing estimates of production function and the true production function. We see that having extra contextual variables does not deteriorate the performance of SCKLS significantly, especially when the input dimension is small and the number of observations is large. Our findings are consistent with the work of Robinson (1988). Since our application data in Section 6 has only two-input, we expect that SCKLS with Z -variables tends not to deteriorate the estimator performance in our application.

Table F.1. RMSE on observation points for experiments with/without Z -variable

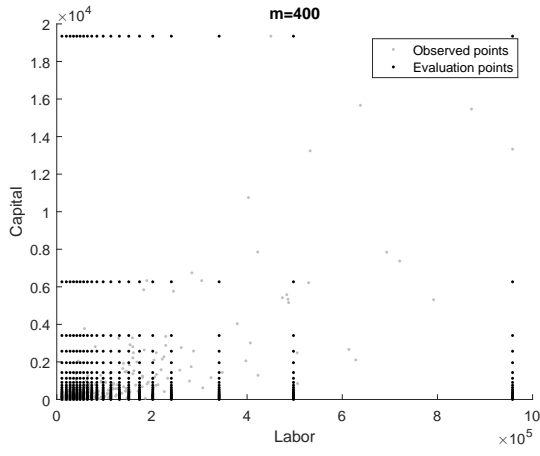
Number of observations		Average of RMSE on observation points				
		100	200	300	400	500
2-input	SCKLS-Z	0.224	0.212	0.239	0.160	0.146
	SCKLS	0.210	0.188	0.170	0.139	0.140
3-input	SCKLS-Z	0.404	0.235	0.261	0.197	0.196
	SCKLS	0.242	0.206	0.215	0.202	0.188
4-input	SCKLS-Z	0.462	0.376	0.332	0.217	0.239
	SCKLS	0.247	0.231	0.202	0.202	0.198

Table F.2. RMSE on evaluation points for experiments with/without Z -variable

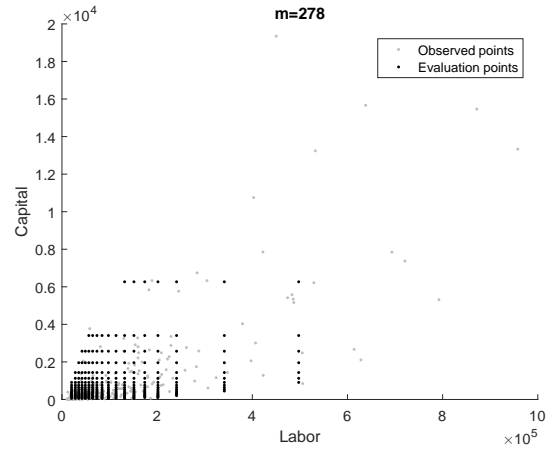
Number of observations		Average of RMSE on evaluation points				
		100	200	300	400	500
2-input	SCKLS-Z	0.245	0.234	0.256	0.172	0.166
	SCKLS	0.230	0.205	0.194	0.154	0.157
3-input	SCKLS-Z	0.496	0.348	0.377	0.271	0.286
	SCKLS	0.316	0.296	0.309	0.271	0.261
4-input	SCKLS-Z	0.648	0.599	0.498	0.397	0.435
	SCKLS	0.385	0.381	0.341	0.350	0.336

G Details on the application to the Chilean manufacturing data

In section 6, we applied the SCKLS estimator to the Chilean manufacturing data to estimate a production function for plastic (2520) and wood (2010) industries. Here we provide the detailed specification of the SCKLS estimator applied to the real data. Since the application data is skewed as shown in Table 6, we use non-uniform grid of evaluation points and limit evaluation points to be inside the convex hull of $\{\mathbf{X}_j\}_{j=1}^n$. Figure G.1 and Figure G.2 show how we set the evaluation points in our application. Originally we set the number of evaluation points is $m = 400$, but after deleting ones which lie outside of the convex hull of $\{\mathbf{X}_j\}_{j=1}^n$, the number is $m \approx 270$ for both industries.

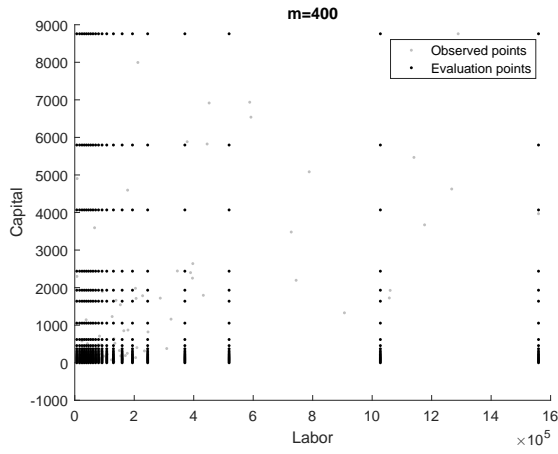


(a) Before deletion

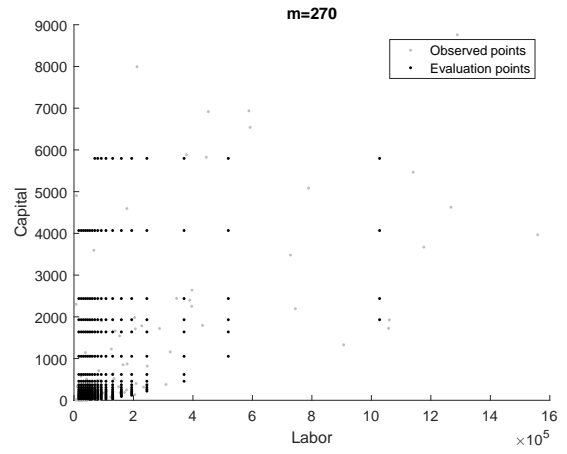


(b) After deletion

Figure G.1. Proposed evaluation points with Plastic industry (2520)



(a) Before deletion



(b) After deletion

Figure G.2. Proposed evaluation points with Wood industry (2010)

References

- Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review* 13(3), 568–598.
- Banker, R. D. and A. Maindiratta (1992). Maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis* 3(4), 401–415.
- Bertsekas, D. (1995). *Nonlinear Programming*. Athena Scientific.
- Chen, X. and Y. J. Qiu (2016). Methods for nonparametric and semiparametric regressions with endogeneity: a gentle guide. Cowles Foundation Discussion Papers 2032, Cowles Foundation for Research in Economics, Yale University.
- Chen, Y. and R. J. Samworth (2016). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society Series B* 78(4), 729–754.
- Chen, Y. and J. A. Wellner (2016). On convex least squares estimation when the truth is linear. *Electronic Journal of Statistics* 10(1), 171–209.
- Dantzig, G., R. Fulkerson, and S. Johnson (1954). Solution of a large-scale traveling-salesman problem. *Journal of the operations research society of America* 2(4), 393–410.
- Dantzig, G. B., D. R. Fulkerson, and S. M. Johnson (1959). On a linear-programming, combinatorial approach to the traveling-salesman problem. *Operations Research* 7(1), 58–66.
- Du, P., C. F. Parmeter, and J. S. Racine (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica* 23(3), 1347–1371.
- Fan, Y. and E. Guerre (2016). Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation. In *Essays in Honor of Aman Ullah*, pp. 489–537. Emerald.
- Frisch, R. (1964). *Theory of production*. Springer.

- Ghosal, P. and B. Sen (2016). On univariate convex regression. arXiv preprint arXiv:1608.04167.
- Grenander, U. (1981). *Abstract Inference*. John Wiley & Sons.
- Hall, P. and L.-S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics* 29(3), 624–647.
- Johnson, A. L. and T. Kuosmanen (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent stonezsd method. *Journal of productivity analysis* 36(2), 219–230.
- Johnson, A. L. and T. Kuosmanen (2012). One-stage and two-stage dea estimation of the effects of contextual variables. *European Journal of Operational Research* 220(2), 559–570.
- Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *The Econometrics Journal* 11(2), 308–325.
- Lee, C.-Y., A. L. Johnson, E. Moreno-Centeno, and T. Kuosmanen (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research* 227(2), 391–400.
- Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.
- Li, Z., G. Liu, and Q. Li (2016). Nonparametric knn estimation with monotone constraints. Working paper.
- Lim, E. and P. W. Glynn (2012). Consistency of multidimensional convex regression. *Operations Research* 60(1), 196–208.
- Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Jounral of Time Series Analysis* 17(6), 571–599.
- Olesen, O. B. and J. Ruggiero (2014). Maintaining the regular ultra passum law in data envelopment analysis. *European Journal of Operational Research* 235(3), 798–809.

- Racine, J. S. (2016). Local polynomial derivative estimation: Analytic or taylor? In *Essays in Honor of Aman Ullah*, pp. 617–633. Emerald.
- Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics* 33(2), 659–680.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica* 56(4), 931–954.
- Sarath, B. and A. Maindiratta (1997). On the consistency of maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis* 8(3), 239–246.
- Sen, B. and M. Meyer (2017). Testing against a linear regression model using ideas from shape-restricted estimation. *Journal of the Royal Statistical Society Series B* 2(79), 423–448.
- van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.
- Varian, H. R. (1984). The nonparametric approach to production analysis. *Econometrica* 52(3), 579–597.