Daisuke Yagi, Yining Chen, Andrew L. Johnson and Timo Kuosmanen

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

Article (Accepted version) (Refereed)

Original citation:
Yagi, Daisuke and Chen, Yining and Johnson, Andrew L. and Kuosmanen, Timo (2018) Shape constrained kernel-weighted least squares: Estimating production functions for Chilean manufacturing industries. Journal of Business and Economic Statistics. 0-0. ISSN 0735-0015
DOI: 10.1080/07350015.2018.1431128

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

Daisuke Yagi[1], Yining Chen[2], Andrew L. Johnson[1,3] and Timo Kuosmanen[4]

[1]Texas A&M University
[2]London School of Economics and Political Science
[3]Osaka University
[4]Aalto University

January 19, 2018

Abstract

In this paper we examine a novel way of imposing shape constraints on a local polynomial kernel estimator. The proposed approach is referred to as Shape Constrained Kernel-weighted Least Squares (SCKLS). We prove uniform consistency of the SCKLS estimator with monotonicity and convexity/concavity constraints and establish its convergence rate. In addition, we propose a test to validate whether shape constraints are correctly specified. The competitiveness of SCKLS is shown in a comprehensive simulation study. Finally, we analyze Chilean manufacturing data using the SCKLS estimator and quantify production in the plastics and wood industries. The results show that exporting firms have significantly higher productivity.

Keywords: Local Polynomials, Kernel Estimation, Multivariate Convex Regression, Nonparametric regression, Shape Constraints.

*We thank two anonymous reviewers and the Associate Editor for providing useful suggestions that helped improve this manuscript. We also thank Chris Parmeter, Jeff Racine and Qi Li for their helpful comments.

1 Introduction

Nonparametric regression methods, such as the local linear (LL) estimator, avoid functional form misspecification. To model production with a production or a cost function, the flexible nature of nonparametric methods can cause difficulties in interpreting the results. Fortunately, microeconomic theory provides additional structure in the form of shape constraints. Recently several nonparametric shape constrained estimators have been proposed that combine the advantage of avoiding parametric functional specification with improved small sample performance relative to unconstrained nonparametric estimators. Nevertheless, the existing methods have limitations regarding either estimation performance or computational feasibility. In this paper, we propose a new estimator that imposes shape restrictions on local kernel weighting methods. By combining local averaging with shape constrained estimation, we improve finite sample performance by avoiding overfitting.

Work on shape-constrained regression first started in the 1950s with Hildreth (1954), who studied the univariate regressor case with a least squares objective subject to monotonicity and concavity/convexity constraints. See also Brunk (1955) and Grenander (1956) for alternative shape constrained estimators. Under the concavity/convexity constraint, properties such as consistency, rate of convergence, and asymptotic distribution have been shown by Hanson and Pledger (1976), Mammen (1991), and Groeneboom et al. (2001), respectively. In the multivariate case, Kuosmanen (2008) developed the characterization of the least squares estimator subject to concavity/convexity and monotonicity constraints, which we will refer to as Convex Nonparametric Least Squares (CNLS) throughout this paper. Furthermore, consistency of the least squares estimator was shown independently by Seijo and Sen (2011) and Lim and Glynn (2012).

Regarding the nonparametric estimation implemented using kernel based methods, Birke and Dette (2007), Carroll et al. (2011), and Hall and Huang (2001) investigated the univariate case and proposed smooth estimators that can impose derivative-based constraints including monotonicity and concavity/convexity. Du et al. (2013) proposed Constrained Weighted Bootstrap (CWB) by generalizing Hall and Huang's method to the multivariate regression setting. Beresteanu (2007) developed a similar type of estimator

but for use with spline based estimators. Finally, we mention the work of Li et al. (2016), which extended Hall and Huang's method to use the k-nearest neighbor approach subject to the monotonicity constraint.

In this paper, *Shape Constrained Kernel-weighted Least Squares* (SCKLS) estimator is described, which optimizes a local polynomial kernel criterion while estimating a multivariate regression function with shape constraints. Under the monotonicity and convex/concavity constraints, we prove uniform consistency and establish the convergence rate of the SCKLS estimator. Kuosmanen (2008), Seijo and Sen (2011) and Lim and Glynn (2012) emphasize the potential advantage that CNLS does not require the selection of tuning parameters. Our proposed SCKLS estimator sheds further light on this issue: in the SCKLS framework, CNLS can be seen as the zero bandwidth estimator; we argue that, compared to unrestricted kernel methods, the SCKLS estimator is relatively robust to the bandwidth selected and is able to alleviate well-known issues such as boundary inconsistency faced by the CNLS estimator.

Note that with n observations, CNLS imposes $O(n^2)$ concavity/convexity constraints, which can lead to computational difficulties. The number of constraints and the number of variables in the SCKLS estimator do not depend on the number of observations, but rather the number of evaluation points which is arbitrarily defined by the modeler, thereby bring the computational complexity of the estimator largely under control of the modeler. In this paper, we implement an iterative algorithm that reduces the number of constraints by building on the ideas in Lee et al. (2013) to further improve the computational performance. We then validate the performance of the SCKLS estimator via Monte Carlo simulations. For a variety of parameter settings, we find performance of SCKLS to be better or at least competitive with CNLS, CWB, and the local linear estimators. We provide the first simulation study of CWB with global concavity constraints. We also investigate the use of variable bandwidth methods that are a function of the data density [1] and propose variants of a uniform grid as practical ways to further improve the performance of SCKLS.

Crucially, we also investigate the behavior of SCKLS when the shape constraints are misspecified and propose a hypothesis test to validate the shape constraints imposed. Hav-

[1] A variable bandwidth method allows the bandwidth associated with a particular regressor to vary with the density of the data.

ing a test that validates the shape constraints is critical because otherwise our estimation procedure would lead to inconsistent estimates.

Finally, we apply the SCKLS estimator empirically on Chilean manufacturing data from the Chilean Annual Industrial Survey. The estimation results provide a concise description of the supply-side of the Chilean plastic and wood industries as we report marginal productivity, marginal rate of substitution and most productive scale size. We also investigate the impact of exporting on productivity by including additional predictors of output in a semi-parametric model. We find that exporting correlates with higher productivity, thus supporting international trade theories that high productivity firms are more likely to compete in international markets.

Our focus on production functions guides our selection of the polynomial function used in estimation, the data generation processes (DGP) in the Monte Carlo simulations. For the application analyzing the Chilean manufacturing data, we are interested in monotonic and concave shape constraints and use a local linear kernel function. These assumptions are motivated by standard economic theory for production functions (Varian, 1984). However, the methods proposed in the paper are general and applicable for other applications with higher order polynomial functions or alternative shape restrictions, as discussed in Appendix A.

The remainder of this paper is as follows. Section 2 describes the model framework and presents our estimator, SCKLS. Section 3 contains the statistical properties of the estimator, and Section 4 discusses the behavior of SCKLS under misspecification, as well as a test for concavity and monotonicity. Monte Carlo simulation results under several different experimental settings are shown in Section 5. Section 6 applies the SCKLS estimator to estimate a production function for both the Chilean plastics and wood industries. Section 7 concludes and suggests future research directions. Appendix A provides extensions to SCKLS and a comparison to CNLS and CWB. Appendix B contains all the technical proofs and Appendix C describes a test for affinity. Appendix D states the details of the iterative algorithm for SCKLS, and Appendix E presents a more extensive set of simulation results. Appendix F describes the details of the partially linear model, and Appendix G gives further details about the application to the Chilean manufacturing data.

2 Model Framework and Methodology

2.1 Model

Suppose we observe n pairs of input and output data, $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$, where for every $j = 1, \ldots, n$, $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jd})' \in \mathbb{R}^d$ is a d-dimensional input vector, and $y_j \in \mathbb{R}$ is an output. Consider the following regression model

$$y_j = g_0(\boldsymbol{X}_j) + \epsilon_j, \quad \text{for } j = 1, \ldots, n,$$

where ϵ_j is a random variable satisfying $E(\epsilon_j | \boldsymbol{X}_j) = 0$. Assume that the regression function $g_0 : \mathbb{R}^d \to \mathbb{R}$ belongs to a class of functions, G, that satisfies certain shape restrictions. Here our estimator can impose any shape restriction that can be modeled as a lower or upper bound on a derivative. Examples are supermodularity, convexity, monotonicity, and quasi-convexity. For purposes of concreteness, and in view of the application to production functions, we focus on imposing monotonicity and global convexity/concavity, specifically, g_0 is concave if:

$$\lambda g_0(\boldsymbol{x_1}) + (1-\lambda) g_0(\boldsymbol{x_2}) \leq g_0(\lambda \boldsymbol{x_1} + (1-\lambda)\boldsymbol{x_2}), \qquad \forall \boldsymbol{x_1}, \boldsymbol{x_2} \in \mathbb{R}^d \text{ and } \forall \lambda \in [0,1]$$

Furthermore, saying g_0 is monotonically increasing means that

$$\text{if } \boldsymbol{x_1} \leq \boldsymbol{x_2}, \text{ then } g_0(\boldsymbol{x_1}) \leq g_0(\boldsymbol{x_2}),$$

where the inequality of $\boldsymbol{x_1} \leq \boldsymbol{x_2}$ means that every component of $\boldsymbol{x_2}$ is greater than or equal to the corresponding component of $\boldsymbol{x_1}$. Here we denote G_2 as the set of functions satisfying these constraints.

2.2 Shape Constrained Kernel-weighted Least Squares (SCKLS) with Local Linear

Given observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$, we state the (multivariate) local linear kernel estimator developed by Stone (1977) and Cleveland (1979) as

$$\min_{a,\boldsymbol{b}} \sum_{j=1}^{n} (y_j - a - (\boldsymbol{X}_j - \boldsymbol{x})'\boldsymbol{b})^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}}{\boldsymbol{h}}\right), \tag{1}$$

where a is a functional estimate, and $\boldsymbol{b}$ is an estimate of the slope of the function at $\boldsymbol{x}$ with $\boldsymbol{x}$ being an arbitrary point in the input space, $K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}}{\boldsymbol{h}}\right)$ denotes a product kernel, and $\boldsymbol{h}$ is a vector of bandwidths (see Racine and Li (2004) for more detail). We note that the objective function uses kernel weights, so more weight is given to the observations that are closer to the point $\boldsymbol{x}$.

We introduce a set of m points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, for evaluating constraints, which we call evaluation points, and impose shape constraints on the local linear kernel estimator. In the spirit of local linear kernel estimator, we define Shape Constrained Kernel-weighted Least Squares (SCKLS) estimator, for the case of monotonicity and concavity, to be the function $\hat{g}_n : \mathbb{R}^d \to \mathbb{R}$ such that

$$\hat{g}_n(\boldsymbol{x}; \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \min_{i \in \{1,\ldots,m\}} \left\{ \hat{a}_i + (\boldsymbol{x} - \boldsymbol{x}_i)'\hat{\boldsymbol{b}}_i \right\} \tag{2}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$, where $\hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_m)'$ and $\hat{\boldsymbol{b}} = (\hat{\boldsymbol{b}}_1', \ldots, \hat{\boldsymbol{b}}_m')'$ are the solutions to the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{a},\boldsymbol{b}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} (y_j - a_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)'\boldsymbol{b}_i)^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right) \\ \text{subject to} \quad & a_i - a_l \geq \boldsymbol{b}_i'(\boldsymbol{x}_i - \boldsymbol{x}_l), \qquad i, l = 1, \ldots, m \\ & \boldsymbol{b}_i \geq 0, \qquad i = 1, \ldots, m. \end{aligned} \tag{3}$$

The first set of constraints in (3) imposes concavity and the second set of constraints imposes non-negativity of $\boldsymbol{b}_i$ at each evaluation point $\boldsymbol{x}_i$. For more details see Kuosmanen

(2008). Note that (2) implies the functional estimate is constructed by taking the minimum of linear interpolations between the evaluation points. This makes SCKLS a globally shape constrained function although it is a non-smooth piece-wise linear function.

The SCKLS estimator requires the user to specify the number and the locations of the evaluation points. A standard method for determining the location of evaluation points, $\{\boldsymbol{x}_i\}_{i=1}^m$, is to construct a uniform grid, where each dimension is divided using equal spacing. However, we can address the skewness of input variable distributions common in manufacturing survey data by using a non-uniform grid method, specifically percentile gridding, to specify evaluation points.

Alternatively, we can deal with the input skewness by applying the k-nearest neighbor (k-NN) approach, Li et al. (2016). The k-NN approach uses a smaller bandwidth in dense data regions and a larger bandwidth when the data is sparse. The analysis in Section 6 uses both a percentile grid and k-NN approach to define the kernel function. For details of these extensions, see Appendix A.

As the density of the evaluation points increases, the estimated function potentially has more hyperplane components and is more flexible; however, the computation time typically increases. If a smooth functional estimate is preferred, see Nesterov (2005) and Mazumder et al. (2015), where methods for smoothing are provided. In practice, we propose to select the bandwidth vector $\boldsymbol{h}$ via the leave-one-out cross-validation based on the unconstrained estimator. See Section 5 for the details.

Appendix A proposes several alternative implementations of the SCKLS estimator: (1) SCKLS with Local Polynomial approximation, (2) a k-nearest neighbor (k-NN) approach and (3) non-uniform grid method.

3 Theoretical Properties of SCKLS

For mathematical concreteness, we next consider the statistical properties of SCKLS under monotonicity and concavity constraints. Recall that G_2 is the class of functions which are monotonically increasing and globally concave, and g_0 is the truth to be estimated from n pairs of observations. We make the following assumptions:

Assumption 1.

(i) $\{\boldsymbol{X}_j, y_j\}_{j=1}^{\infty}$ *are a sequence of i.i.d. random variables with* $y_j = g_0(\boldsymbol{X}_j) + \epsilon_j$.

(ii) $g_0 \in G_2$ *and is twice-differentiable.*

(iii) $\boldsymbol{X}_j$ *follows a distribution with continuous density function* f *and support* $\boldsymbol{S}$. *Here* $\boldsymbol{S}$ *is a convex, non-degenerate and compact subset of* $\mathbb{R}^d$. *Moreover,*

$$\min_{\boldsymbol{x} \in \boldsymbol{S}} f(\boldsymbol{x}) > 0.$$

(iv) The conditional probability density function of ϵ_j, *given* $\boldsymbol{X}_j$, *denoted as* $p(e|\boldsymbol{x})$, *is continuous with respect to both* e *and* $\boldsymbol{x}$, *with the mean function*

$$\mu(\cdot) = E(\epsilon_j | \boldsymbol{X}_j = \cdot) = 0$$

and the variance function

$$\sigma^2(\cdot) = \mathrm{Var}(\epsilon_j | \boldsymbol{X}_j = \cdot)$$

bounded away from 0 and continuous over $\boldsymbol{S}$. *Moreover,* $\sup_{\boldsymbol{x} \in \boldsymbol{S}} E\left(\epsilon_j^4 \middle| \boldsymbol{X}_j = \boldsymbol{x}\right) < \infty$.

(v) $K(\cdot)$ *is a non-negative, Lipschitz second order kernel with a compact and convex support. For simplicity, we set the bandwidth associated with each explanatory variable,* h_k, *for* $k = 1, \ldots, d$, *to be* $h_1 = \cdots = h_d = h$.

(vi) $h = O(n^{-1/(4+d)})$ *as* $n \to \infty$.

Here (i) states that the data are i.i.d.; (ii) says that the constraints we impose on the SCKLS estimator are satisfied by the true function; (iii) makes a further assumption on the distribution of the covariates; (iv) states that the noise can be heteroscedastic in certain ways, but requires the change in the variance to be smooth; (v) is rather standard in local polynomial estimation to facilitate the theoretical analysis; and (vi) assures the bandwidths become sufficiently small as $n \to \infty$ so that both the bias and the variance from local averaging go to zero. For details of the consistency of local linear estimator

and a discussion of some of these conditions, see Masry (1996), Li and Racine (2007) and Fan and Guerre (2016).

We consider two scenarios: let the number of evaluation points (denoted by m) grow with n, or fix the number of evaluation points a priori. For simplicity, we also assume that the evaluation points are drawn independent of $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$.

Assumption 2.

(i) The number of evaluation points $m \to \infty$ as $n \to \infty$. For simplicity, we assume that the empirical distribution of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ converges to a distribution Q that has support $\boldsymbol{S}$ (i.e. as defined in Assumption 1(iv))) and a continuous differentiable density function $q : \boldsymbol{S} \to \mathbb{R}$ satisfying $\min_{\boldsymbol{x} \in \boldsymbol{S}} q(\boldsymbol{x}) > 0$.

(ii) The number of evaluation points m is fixed. All the evaluation points lie in the interior of $\boldsymbol{S}$. Moreover,

$$\frac{\sup_{\boldsymbol{x} \in \boldsymbol{S}} \min_{i=1,\ldots,m} \|\boldsymbol{x} - \boldsymbol{x}_i\|}{\min_{i \neq j; i,j \in \{1,\ldots,m\}} \|\boldsymbol{x}_j - \boldsymbol{x}_i\|} \leq \kappa$$

for some $\kappa \geq 1$ (i.e. $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ are reasonably well spread across $\boldsymbol{S}$).

Our main results are summarized below. A short discussion on our proof strategy and the proofs are available in Appendix B.

Theorem 1. *Suppose that Assumption 1(i)-1(vi) and Assumption 2(i) or 2(ii) hold. Then,*

$$\frac{1}{m} \sum_{i=1}^{m} \{\hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)\}^2 = O(n^{-4/(4+d)} \log n)$$

.

Theorem 2.

1. ***(The case of an increasing** m**)** *Suppose that Assumption 1(i)-1(vi) and Assumption 2(i) hold. Let $\boldsymbol{C}$ be any fixed closed set that belongs to the interior of $\boldsymbol{S}$. Then with probability one, as $n \to \infty$, the SCKLS estimator satisfies*

$$\sup_{\boldsymbol{x} \in \boldsymbol{C}} \left| \hat{g}_n(\boldsymbol{x}) - g_0(\boldsymbol{x}) \right| \to 0.$$

2. ***(The case of a fixed** m**)** *Suppose that Assumption 1(i)-1(vi) and Assumption 2(ii) hold. Then, as* $n \to \infty$*, with probability one, the estimates from SCKLS satisfy*

$$\hat{a}_i \to g_0(\boldsymbol{x}_i) \quad \text{and} \quad \hat{\boldsymbol{b}}_i \to \frac{\partial g_0}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)$$

for all $i = 1, \ldots, m$*.*

Note that this convergence rate is nearly optimal (differing only by a factor of $\log n$). However, in the above, we only manage to show that the SCKLS estimator converges at the evaluation points or in the interior of the domain. It is known that shape-constrained estimators tend to suffer from bad boundary behaviors. For instance, the quantity $\sup_{\boldsymbol{S}} \left|\hat{g}_n^{CNLS}(\boldsymbol{x}) - g_0(\boldsymbol{x})\right|$ does *not* converge to zero in probability, where $\hat{g}_n^{CNLS}$ is the CNLS estimator. Though for SCKLS, if we let the number of evaluation points, m, grow at a rate slower than n, we argue that we can both alleviate the boundary inconsistency and improve the computational efficiency.

Assumption 3. *The number of evaluation points* $m = o(n^{2/(4+d)}/\log n)$ *as* $n \to \infty$*.*

Theorem 3. *Suppose that Assumption 1(i)-1(vi), Assumption 2(i) and Assumption 3 hold. Then, with probability one, as* $n \to \infty$*, the SCKLS estimator satisfies*

$$\sup_{\boldsymbol{x} \in \boldsymbol{S}} \left|\hat{g}_n(\boldsymbol{x}) - g_0(\boldsymbol{x})\right| \to 0.$$

We also note that CNLS can be viewed as a special case of SCKLS when we let the set of evaluation points be $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and the bandwidth vector $\|\boldsymbol{h}\| \to \boldsymbol{0}$. See Appendix A for the proof of the relationship between CNLS and SCKLS, together with more discussions on the relationship between SCKLS and alternative shape constrained estimators such as CWB.

4 Shape Misspecification: Theory and Testing

4.1 Misspecification of the shape restrictions

So far we have assumed in our estimation procedures that $g_0 \in G_2$, where G_2 is the class of functions which are monotonically increasing and globally concave. To understand the behavior of SCKLS, we are interested in its performance when $g_0 \notin G_2$.

Let Q be a distribution on $\boldsymbol{S}$ (as in Assumption 2(i)) and define $g^* : \boldsymbol{S} \to \mathbb{R}$ as

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

The existence and Q-uniqueness of g_0^* follows from the well-known results about the projection onto a cone in the Hilbert space. When $g_0 \in G_2$, it is easy to check that $g_0^* = g_0$. See also Lim and Glynn (2012). The following result can be viewed as a generalization of Theorem 2.

Theorem 4.

Suppose that Assumption 1(i), 1(iii)-1(vi) and Assumption 2(i) hold. Furthermore, suppose that g_0 is twice-differentiable. Let $\boldsymbol{C}$ be any compact set that belongs to the interior of $\boldsymbol{S}$. Then with probability one, as $n \to \infty$, the SCKLS estimator satisfies

$$\sup_{\boldsymbol{x} \in \boldsymbol{C}} \left| \hat{g}_n(\boldsymbol{x}) - g_0^*(\boldsymbol{x}) \right| \to 0.$$

Theorem 4 assures us that the SCKLS estimator converges uniformly on a compact set to the function g_0^* that is closest in L^2 distance to the true function g_0 for which our estimator is misspecified. Consequently, as long as g_0 is not too far away from G_2, our estimator can still be used as a reasonable approximation to the truth, especially when the sample size is moderate. See Appendix E for a numerical demonstration.

4.2 Hypothesis Testing for the Shape

Admittedly, the SCKLS estimator can be inappropriate if the shape constraints are not fulfilled by g_0. Thus, we propose a procedure based on the SCKLS estimators for testing

$$H_0: \ \{g_0 : \boldsymbol{S} \to \mathbb{R}\} \in G_2 \quad \text{against} \quad H_1: \ \{g_0 : \boldsymbol{S} \to \mathbb{R}\} \notin G_2.$$

Denote by

$$\tilde{r}^2\Big(\{\boldsymbol{X}_j, y_j\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big) = \min_{\boldsymbol{a},\boldsymbol{b}} \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)'\boldsymbol{b}_i)^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right);$$

the value of the objective function that is minimized by the local linear kernel estimator. And denote by

$$\hat{r}^2\Big(\{\boldsymbol{X}_j, y_j\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big) = \min_{\boldsymbol{a},\boldsymbol{b}} \sum_{i=1}^m \sum_{j=1}^n (y_j - a_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)'\boldsymbol{b}_i)^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right),$$
$$\text{subject to} \quad a_i - a_l \geq \boldsymbol{b}_i'(\boldsymbol{x}_i - \boldsymbol{x}_l) \text{ and } \boldsymbol{b}_i \geq 0, \ i, l = 1, \ldots, m.$$

Here $\hat{r}^2(\cdot, \cdot)$ is the value of the objective function that is minimized by SCKLS.

We focus on the test statistic

$$T_n := T\Big(\{\boldsymbol{X}_j, y_j\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big) = \Big[\frac{1}{mnh^d}\Big\{\hat{r}^2\Big(\{\boldsymbol{X}_j, y_j\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big) - \tilde{r}^2\Big(\{\boldsymbol{X}_j, y_j\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big)\Big\}\Big]^{1/2},$$

which is a re-scaled version of the difference between the values of the same objective function (with the same bandwidth $\boldsymbol{h}$), optimized either with or without the shape constraints. Intuitively, the value of this statistic should be small if $g_0 \in G_2$. This statistic can also be viewed as a smoothed and re-scaled version of the goodness-of-fit statistic.

Here we focus on the boundary case when g_0 is constant (i.e. $g_0 = 0$) because it is hardest to evaluate the null hypothesis when g_0 is both non-increasing and non-decreasing and both concave and convex, intuitively and theoretically and it allows us to control the size of our test statistic. Since the noise here might be non-homogeneous, we use the wild bootstrap to approximate the distribution of the test statistic under H_0. See Wu (1986),

Liu (1988), Mammen (1993) and Davidson and Flachaire (2008) for an overview of the wild bootstrap procedure.

Our testing procedure has three steps:

1. Estimate the error at each $\boldsymbol{X}_j$ by $\tilde{\epsilon_j} = y_j - \tilde{g}_n(\boldsymbol{X}_j)$ for $j = 1, \ldots, n$, where $\tilde{g}$ is the unconstrained local linear estimator with kernel and bandwidth satisfying Assumptions 1(v)–(vi).

2. The wild bootstrap method is used to construct a critical region for T_n. Let B be the number of Monte Carlo iterations. For every $k = 1, \ldots, B$, let $\boldsymbol{u}_k = (u_{1k}, \ldots, u_{nk})'$ be a random vector with components sampled independently from the Rademacher distribution, i.e. $P(u_{jk} = 1) = P(u_{jk} = -1) = 0.5$. Furthermore, let $y_{jk} = u_{jk}\ \tilde{\epsilon}_j$. Then, the wild bootstrap test statistic is

$$T_{nk} = T\Big(\{\boldsymbol{X}_j, y_{jk}\}_{j=1}^n, \{\boldsymbol{x}_i\}_{i=1}^m\Big).$$

3. Define the Monte Carlo p-value as[2]

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

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

A few remarks are in order.

First, here we conveniently implemented the simplest wild bootstrap scheme to simplify our analysis, in line with the work of Davidson and Flachaire (2008). Instead of imposing the Rademacher distribution on u_{kj}, we can also use any distribution with zero-mean and unit-variance. One popular choice suggested by Mammen (1993) is

$$u_{jk} = \begin{cases} -\frac{\sqrt{5}-1}{2} & \text{with probability } \frac{5+\sqrt{5}}{10} \\ \frac{\sqrt{5}+1}{2} & \text{with probability } \frac{5-\sqrt{5}}{10} \end{cases}.$$

[2]Since we underestimate the level of the errors in Step 1 by a factor of roughly $n^{-2/(4+d)}$, for the theoretical development, we address this bias issue by modifying the p-value to be $p_n = \frac{1}{B}\sum_{k=1}^{B} \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}}$, where $\Delta_n = O(n^{-2/(4+d)} \log n)$. Note that if we fix m and pick $h = O(n^{-\eta})$ for $\eta \in (\frac{1}{4+d}, \frac{1}{d})$, then $\Delta_n / T_{nk} = o_p(1)$ as $n \to \infty$, i.e. this correction has a negligible effect. Indeed, our experience suggests that this modification offers little improvement in terms of finite sample performance in our simulation study.

Second, note that the definition of y_{jk} in Step 2 makes this a test of the residuals, i.e., when drawing bootstrap samples, we use $y_{jk} = u_{jk}\,\tilde{\epsilon}_j$ instead of $y_{jk} = \hat{g}_n(\boldsymbol{X}_j) + u_{jk}\,\tilde{\epsilon}_j$. From this perspective, our test is similar to the univariate monotonicity test in Hall and Heckman (2000). One reason behind this choice is to avoid the boundary inconsistency of the bootstrap procedure. See Andrews (2000) and Cavaliere et al. (2017) who addressed this issue in a much simpler setup. Generally speaking, testing the null hypothesis becomes harder when g_0 is on the boundary of G_2. In practice, we could use $y_{jk} = \hat{g}_n(\boldsymbol{X}_j) + u_{jk}\,\tilde{\epsilon}_j$ in certain scenarios (e.g. when testing g_0 is a strictly increasing and strictly concave function against $g_0 \notin G_2$), and slight improvements are observed in terms of finite-sample performance.

We now look into the theoretical properties of our procedure under both H_0 and H_1. See Appendix B for the proof.

Theorem 5. *Suppose that Assumptions 1(i),(iii)–(v) and 2(i) hold, and the conditional error distribution (i.e. $\epsilon_j|\boldsymbol{X}_j$) is symmetric. Furthermore, assume that g_0 is continuously twice-differentiable and let $h = O(n^{-\eta})$ for some fixed $\eta \in (\frac{1}{4+d}, \frac{1}{d})$. Let $B := B(n) \to \infty$ as $n \to \infty$. Then, for any given $\alpha \in (0,1)$,*

- *Type I error: for any $g_0 \in G_2$, $\limsup_{n\to\infty} P(p_n < \alpha) \leq \alpha$;*
- *Type II error: for any $g_0 \notin G_2$, $\limsup_{n\to\infty} \Big\{1 - P(p_n < \alpha)\Big\} = 0$.*

In addition, if we replace Assumption 2(i) by Assumption 2(ii), the same conclusions hold for sufficiently large m.

See also Section 5 for the finite-sample performance of our test in a simulation study, where we demonstrate that the proposed test controls both Type I and Type II errors reasonably well. Additionally, Appendix C describes our procedure for testing affinity using SCKLS.

5 Simulation study

5.1 Numerical experiments on estimation

5.1.1 The setup

We now examine the finite sample performance and robustness of the proposed estimator through Monte Carlo simulations. We run our experiments on a computer with Intel Core2 Quad CPU 3.00 GHz and 8GB RAM. We compare the performance of SCKLS is compared with that of CNLS and LL. See Appendix E for a comparisons of SCKLS with CWB. For the SCKLS and the CNLS estimator, we solve the quadratic programming problems with MATLAB using the built-in quadratic programming solver, `quadprog`. We run two sets of experiments varying the number of observations (n), the number of evaluation points (m), and the number of the inputs (d). We also run additional experiments to show the robust performance of the SCKLS estimator under alternative conditions. See Appendix E for the results.

We measure the estimator's performance using Root Mean Squared Errors (RMSE) based on two criteria: the distance from the estimated function to the true function measured 1) at the observed points and 2) at the evaluation points constructed on an uniform grid , respectively. As CNLS estimates hyperplanes at observation points, we use linear interpolation to obtain the RMSE of CNLS[3]. We replicate each scenario 10 times and report the average and standard deviation.

5.1.2 Choosing of the tuning parameters

For the SCKLS estimator, we use the Gaussian kernel function $K(\cdot)$ and leave-one-out cross-validation (LOOCV) for bandwidth selection. LOOCV is a data-driven method, and has been shown to perform well for unconstrained kernel estimators such as local linear (Stone, 1977). We apply LOOCV procedure on unconstrained estimates (i.e. local linear) to select the bandwidth for SCKLS to reduce the computational burden and because SCKLS is

[3]The CNLS estimates include the second stage linear programming estimation procedure described in Kuosmanen and Kortelainen (2012) to find the minimum extrapolated production function.

relatively insensitive to the bandwidth choice (see for example Section 5.1.3.1). For further computational improvements, we apply the iterative algorithm described in Appendix D.

5.1.3 Results

5.1.3.1 Fixed number of evaluation points

Experiment 1. We consider a Cobb–Douglas production function with d-inputs and one-output, $g_0(x_1, \ldots, x_d) = \prod_{k=1}^{d} x_k^{\frac{0.8}{d}}$. For each pair $(\boldsymbol{X}_j, y_j)$, each component of the input, $\boldsymbol{X}_{jk}$, is randomly and independently drawn from uniform distribution $unif[1, 10]$, 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 dimensions (2, 3 and 4). The structure and data generation process of Experiment 1 follows Lee et al. (2013). We fix the number of evaluation points at approximately 400 and locate them on a uniform grid.

For this experiment, we compare the following four estimators: SCKLS, CNLS, Local Linear Kernel (LL), and parametric Cobb–Douglas estimator. The latter estimator serves as a baseline because it is correctly specified parametric form. Tables 1 and 2 show for Experiment 1 the RMSE measured on observation points and evaluation points, respectively. The number in parentheses is the standard deviation of RMSE values computed by 10 replications. Note the standard derivations are generally small compared to the parameter estimates, which indicates low variability even after only 10 replications. A more extensive set of results for this experiment is summarized in Appendix E. The SCKLS estimator has the lowest RMSE in most scenarios even when RMSE is measured on observation points (note that the SCKLS estimator imposes the global shape constraints via evaluation points in Equation (3)). Also as expected, the performance of SCKLS estimator improves as the number of observation points increases. Moreover, the SCKLS estimator performs better than the LL estimator particularly in higher dimensional functional estimation. This provides empirical evidence that the shape constraints in SCKLS are helpful in improving the finite sample performance as compared to LL. Note that LL appears to have larger RMSE values on evaluation points which are located in input space regions with sparse observations. This implies that the SCKLS estimator has more robust out-of-sample performance

than the LL estimator due to the shape constraints. We also observe that the performance of the CNLS estimator measured at the evaluation points is worse than that measured at the observations. CNLS often has ill-defined hyperplanes which are very steep/shallow at the edge of the observed data, and this over-fitting leads to poor out-of-sample performance. In contrast, the SCKLS estimator performs similarly for both the observation points and evaluation points, because the construction of the grid that completely covers the observed data makes the SCKLS estimator more robust.

We also conduct simulations with different bandwidths to analyze the sensitivity of each estimator to bandwidths. We compare SCKLS and LL with bandwidth $h \in [0, 10]$ with an increment by 0.01 for the 1-input setting, and we use bandwidth $\boldsymbol{h} \in [0, 5] \times [0, 5]$ with an increment by 0.25 in each coordinate for the 2-input setting. We simulate 100 datasets to compute the RMSE for each bandwidth as well as for the bandwidth via LOOCV. Figure 1 displays the average RMSE of each estimator. The histogram shows the distribution of bandwidths selected by LOOCV. The instances when SCKLS and LL provide the lowest RMSE are shown in light gray and dark gray respectively. For the one-input scenario, the SCKLS estimator performs better than the LL estimator for bandwidth between 0.25 -

Table 1. RMSE on observation points for Experiment 1.

		Average of RMSE on observation points				
Number of observations		100	200	300	400	500
2-input	SCKLS	**0.193**	0.171	0.141	**0.132**	0.118
		(0.053)	(0.047)	(0.032)	(0.029)	(0.017)
	CNLS	0.229	**0.163**	**0.137**	0.138	**0.116**
		(0.042)	(0.037)	(0.010)	(0.027)	(0.016)
	LL	0.212	0.166	0.149	0.152	0.140
		(0.079)	(0.042)	(0.028)	(0.028)	(0.028)
	Cobb–Douglas	0.078	0.075	0.048	0.039	0.043
3-input	SCKLS	**0.230**	**0.187**	**0.183**	**0.152**	**0.165**
		(0.050)	(0.026)	(0.032)	(0.019)	(0.031)
	CNLS	0.294	0.202	0.189	0.173	0.168
		(0.048)	(0.035)	(0.020)	(0.014)	(0.020)
	LL	0.250	0.230	0.235	0.203	0.181
		(0.068)	(0.050)	(0.052)	(0.050)	(0.021)
	Cobb–Douglas	0.104	0.089	0.070	0.047	0.041
4-input	SCKLS	**0.225**	**0.248**	**0.228**	**0.203**	**0.198**
		(0.038)	(0.020)	(0.037)	(0.042)	(0.028)
	CNLS	0.315	0.294	0.246	0.235	0.214
		(0.039)	(0.027)	(0.024)	(0.029)	(0.015)
	LL	0.256	0.297	0.252	0.240	0.226
		(0.044)	(0.057)	(0.056)	(0.060)	(0.038)
	Cobb–Douglas	0.120	0.073	0.091	0.067	0.063

Table 2. RMSE on evaluation points for Experiment 1.

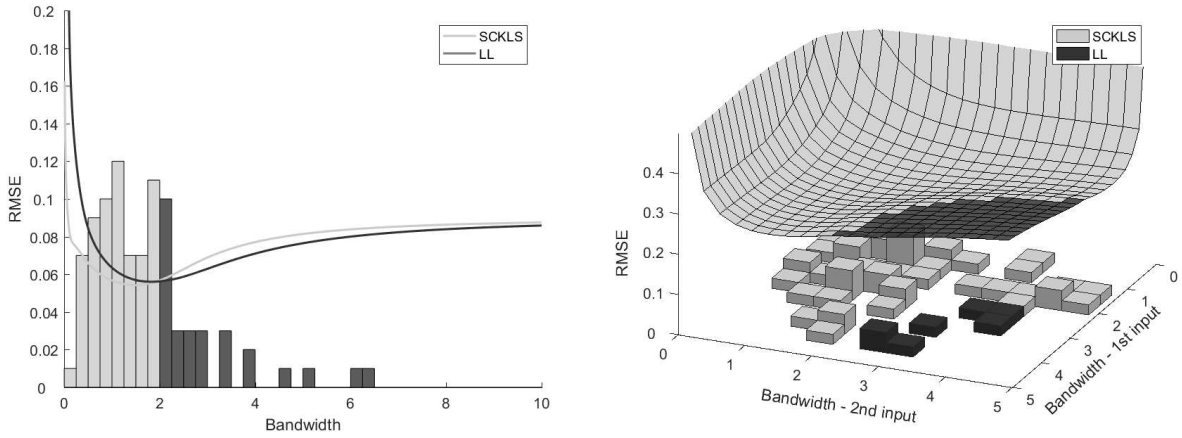
Number of observations		Average of RMSE on evaluation points 100	200	300	400	500
2-input	SCKLS	**0.219**	0.189	**0.150**	**0.147**	**0.128**
		(0.053)	(0.057)	(0.034)	(0.030)	(0.021)
	CNLS	0.350	0.299	0.260	0.284	0.265
		(0.082)	(0.093)	(0.109)	(0.119)	(0.078)
	LL	0.247	**0.182**	0.167	0.171	0.156
		(0.101)	(0.053)	(0.030)	(0.030)	(0.034)
	Cobb–Douglas	0.076	0.076	0.049	0.040	0.043
3-input	SCKLS	**0.283**	**0.231**	**0.238**	**0.213**	**0.215**
		(0.072)	(0.033)	(0.030)	(0.029)	(0.034)
	CNLS	0.529	0.587	0.540	0.589	0.598
		(0.112)	(0.243)	(0.161)	(0.109)	(0.143)
	LL	0.336	0.340	0.360	0.326	0.264
		(0.085)	(0.093)	(0.108)	(0.086)	(0.042)
	Cobb–Douglas	0.116	0.098	0.080	0.052	0.046
4-input	SCKLS	**0.321**	**0.357**	**0.329**	**0.308**	**0.290**
		(0.046)	(0.065)	(0.049)	(0.084)	(0.044)
	CNLS	0.845	0.873	0.901	0.827	0.792
		(0.188)	(0.137)	(0.151)	(0.235)	(0.091)
	LL	0.482	0.527	0.483	0.495	0.445
		(0.115)	(0.125)	(0.146)	(0.153)	(0.074)
	Cobb–Douglas	0.146	0.091	0.115	0.081	0.080

2.25 as shown in (a). For the 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, calculated using the unconstrained estimator, provides bandwidths that work well for the SCKLS estimator. Importantly, the SCKLS estimator does not appear to be very sensitive to the bandwidth selection method since, heuristically, the shape constraints help reduce the variance of the estimator. Finally, we note that similar results can be obtained in experimental settings with lower signal-to-noise level, or with non-uniform input. See Appendix E for more details.

5.1.3.2 Different numbers of evaluation points

Experiment 2. The setting is the same as Experiment 1. However, now we consider 9 different scenarios with different numbers of evaluation points (100, 300 and 500) and input dimensions (2, 3 and 4). We fix the number of observed points at 400.

We show the performance of SCKLS. Table 3 and 4 shows for Experiment 2 the RMSE measured on observations and evaluation points respectively. Both tables show that empirically even if we increase the number of evaluation points, the RMSE value does not



(a) One-input (b) Two-input

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

change significantly. This has important implications for the running time. Specifically, we can reduce the calculation time by using a rough grid without sacrificing too much in terms of RMSE performance of the estimator.

Table 3. RMSE on observation points for Experiment 2.

Number of evaluation points		Average of RMSE on observation points 100	300	500
2-input	SCKLS	0.142	0.141	0.141
3-input	SCKLS	0.198	0.203	0.197
4-input	SCKLS	0.239	0.207	0.206

Table 4. RMSE on evaluation points for Experiment 2.

Number of evaluation points		Average of RMSE on evaluation points 100	300	500
2-input	SCKLS	0.181	0.164	0.158
3-input	SCKLS	0.304	0.267	0.257
4-input	SCKLS	0.383	0.296	0.270

5.2 Numerical experiments on testing the imposed shape

Experiment 3. We test monotonicity and concavity for data generated from the following single-input and single-output DGP:

$$g_0(x) = x^p \tag{4}$$

and

$$g_0(x) = \frac{1}{1 + \exp(-5\log(2x))}. \tag{5}$$

With n observations, for each pair (X_j, y_j), each input, X_j, is randomly and independently drawn from uniform distribution $unif[0, 1]$. In this simulation, we use the following multiplicative noise to validate whether the wild bootstrap can handle non-homogeneous noise.

$$y_j = g_0(X_j) + (X_j + 1) \cdot \epsilon_j,$$

where ϵ_j, is randomly and independently sampled from a normal distribution, $N(0, \sigma^2)$. We use three different DGP scenarios A, B and C. For scenarios A and B, we use function (4) where the exponent parameter p defines whether the function g_0 is an element of the class of functions G_2 or not. We use $p = \{0, 2\}$ for scenarios A and B respectively, where $g_0 \in G_2$ if $p = 0$, and $g_0 \notin G_2$ if $p = 2$ since g_0 is strictly convex. For scenario C, we consider an "S"-shape function defined by (5) which violates both global concavity and convexity. We consider different sample sizes $n = \{100, 300, 500\}$ and standard deviation of the noise $\sigma = \{0.1, 0.2\}$, and perform 500 simulations to compute the rejection rate for each scenario. We assume that we do not know the distribution of the noise in advance and use the wild bootstrap procedure described in Section 4.2 with $B = 200$.

Table 5 shows the rejection rate for each DGP. For high signal-to-noise ratio scenarios ($\sigma = 0.1$), the test works well even with a small sample size. Our test is able to control the Type I error, as illustrated in scenario A. In addition, the Type II error of our test is small for the scenarios B and C where shape constraints are violated by the DGP. Furthermore, for low signal-to-noise ratio scenarios ($\sigma = 0.2$), the rejection rate for scenarios B and C significantly improves when the sample size is increased from 100 to 300. Indeed, for

larger noise scenarios more data is required for the test to have power. Thus, our test seems informative enough to guide users to avoid imposing shape constraints on the data generated from misspecified functions.

Table 5. Rejection rate (%) of the test for monotonicity and concavity

Sample size (n)	DGP Scenario	Power of the Test (α) 0.05	0.01	0.05	0.01
		$\sigma = 0.1$		$\sigma = 0.2$	
100	A (H_0)	5.8	2.0	8.0	2.6
	B (H_1)	98.6	94.6	55.0	36.2
	C (H_1)	98.6	94.4	42.6	24.2
300	A (H_0)	6.8	1.8	6.6	3.0
	B (H_1)	100.0	100.0	92.0	83.2
	C (H_1)	100.0	100.0	97.0	86.8
500	A (H_0)	5.4	1.6	5.6	1.4
	B (H_1)	100.0	100.0	99.4	97.2
	C (H_1)	100.0	100.0	99.8	99.4

6 Application

We apply the proposed method to estimate the production function for two large industries in Chile: plastic (2520) and wood manufacturing (2010) where the values inside the parentheses indicate the CIIU3 industry code. There are some existing studies which analyze the productivity of Chilean data, see for example Pavcnik (2002), who analyzed the effect of trade liberalization on productivity improvements. Other researchers have analyzed the productivity of Chilean manufacturing including Benavente (2006), Alvarez and Görg (2009) and Levinsohn and Petrin (2003). However, the above-cited work use strong parametric assumptions and older data. Most studies use the Cobb–Douglas functional form which restricts the elasticity of substitution to be 1. When diminishing marginal productivity of inputs characterizes the data, the Cobb–Douglas functional form imposes that the most productive scale size is at the origin. We relax the parametric assumptions and estimate a shape constrained production function nonparametrically using data from 2010. We examine the marginal productivity, marginal rate of substitution, and most productive scale size (MPSS) to analyze the structure of the industries. We also investigate how productivity differs between exporting and non-exporting firms, as exporting has become an important source of revenue in Chile[4]. See Appendix G for the details of estimation and comparison across different estimators.

6.1 The census of Chilean manufacturing plants

We use the Chilean Annual Industrial Survey provided by Chile's National Institute of Statistics[5]. The survey covers manufacturing establishments with ten or more employees. We define Capital and Labor as the input variables and Value Added as the output variable of the production function[6]. Capital and Value Added are measured in millions of Chilean

[4]Note that firms' decisions, i.e., selecting labor and capital levels with considerations for productivity levels or whether to export, are potentially endogenous. Solutions to this issue are to instrument or build a structural model based on timing assumptions. Our estimator can be embedded within the estimation procedures such as those described in Ackerberg et al. (2015) to address this issue.

[5]The data are available at `http://www.ine.cl/estadisticas/economicas/manufactura`.

[6]The definition of Labor includes full-time, part-time, and outsourced labors. Capital is defined as a sum of the fixed assets balance such as buildings, machines, vehicles, furniture, and technical software. Value added is computed by subtracting the cost of raw materials and

peso while Labor is measured as the total man-hours per year. We use cross sectional data from the plastic and the wood industries.

Many researchers have found positive effects of exporting for other countries using parametric models. See for instance, De Loecker (2007) and Bernard and Jensen (2004). Here we use SCKLS to relax the parametric assumption for the production function. To capture the effects of exporting, we use a semi-parametric modeling extension of SCKLS. The partially linear model is represented as follows:

$$y_j = \boldsymbol{Z}_j'\boldsymbol{\gamma} + g_0(\boldsymbol{X}_j) + \epsilon_j, \tag{6}$$

where $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2})'$ denotes contextual variables and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$ is the coefficient of contextual variables. We model exporting with two variables: a dummy variable indicating the establishments that are exporting and the share of output being exported. For more details see Appendix F.

Table 6 presents the summary of statistics for each industry by exporter/non-exporter. We find that exporters are typically larger than non-exporter in terms of labor and capital. Input variables are positively skewed, indicating there exist many small and few large establishments. Since SCKLS with variable bandwidth (k-nearest neighbor) and non-uniform grid performed the best in our simulation scenarios with non-uniform input data (as indicated in Appendix E), we use these options. We choose the smoothing parameter k via leave-one-out cross validation. Appendix A explains the details of our implementation of K-NN for the SCKLS estimator.

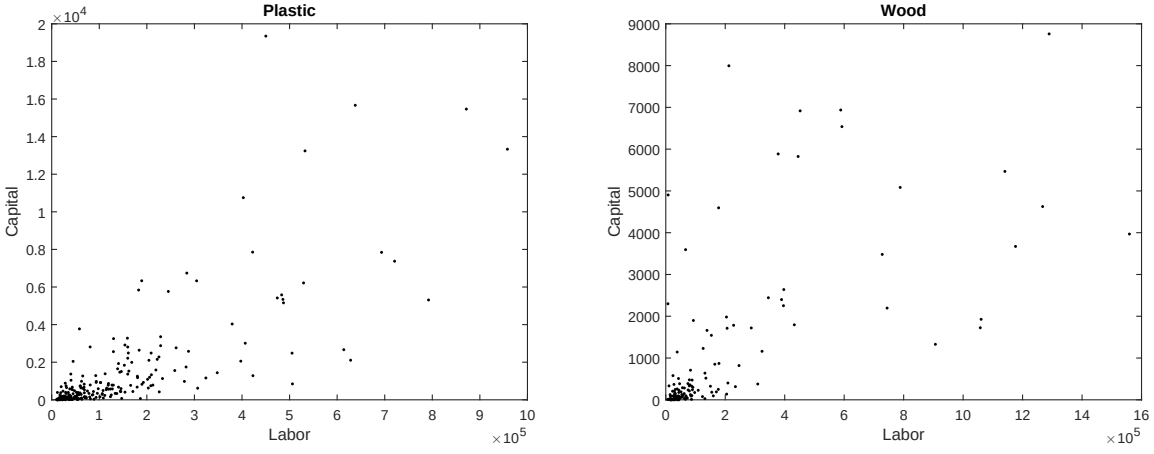
Figure 2 is a plot of labor and capital for each industry and shows input data is sparse for large establishments. Beresteanu (2005) proposed to include shape constraints only for the evaluation points that are close to the observations. Thus, in addition to using a percentile grid of evaluation points, we propose to use the evaluation points that are inside the convex hull of observed input $\{\boldsymbol{X}_j\}_{j=1}^n$. See Appendix G for details.

We begin by testing if the Cobb–Douglas production function is appropriate for our data. We use the hypothesis test for correct parametric specification described in Henderson and Parmeter

intermediate consumption from the total amount produced. Further details are available at `http://www.ine.cl/estadisticas/economicas/manufactura`.

Table 6. Statistics of Chilean manufacturing data.

Plastic (2520)	Non-exporters ($n = 173$)			Exporters ($n = 72$)			
	Labor	Capital (million)	Value Added (million)	Labor	Capital (million)	Value Added (million)	Share of Exports
mean	92155	725.85	546.93	240890	2859	1733.9	0.147
median	55220	258.41	247.05	180330	1329.1	1054.9	0.0524
std	106530	1574	1068.1	212480	3840.2	1678.8	0.201
skewness	3.301	5.2052	5.9214	1.3681	2.4594	1.0678	-0.303
Wood (2010)	Non-exporters ($n = 97$)			Exporters ($n = 35$)			
	Labor	Capital	Value Added	Labor	Capital	Value Added	Share of Exports
mean	76561	364.93	334.83	501470	3063.4	4524.1	0.542
median	44087	109.48	115.39	378000	2195.4	2673.5	0.648
std	78057	702.35	555.87	436100	2510.3	4466.3	0.355
skewness	2.243	3.5155	3.432	0.81454	0.63943	1.0556	-0.303



(a) Plastic (2520) (b) Wood (2010)

Figure 2. Labor and Capital of each industry.

(2015)[7]. The resulting p-value is 0.092 for the plastic industry and 0.007 for the wood industry, respectively. Therefore, the Cobb–Douglas parametric specification is likely to be wrong, particularly applied to the wood industry.

Next, we apply the test proposed in Section 4.2 to determine if imposing global concavity and monotonicity shape constraints is appropriate. We estimate a p-value of 0.302 for the plastic industry and 0.841 for the wood industry, respectively. For both industries, the estimated p-value is not small enough to reject H_0, which means that the observed data is likely to satisfy the shape constraints imposed.

6.2 Estimated production function and interpretation

We estimate a semi-parametric model with a nonparametric shape constrained production function, a linear model for exporting share of sales, and a dummy variable for exporting. Table 7 shows the goodness of fit (R^2) of the production function: 71.1% of variance is explained in the plastic industry while 43.8% of variance is explained in the wood industry.

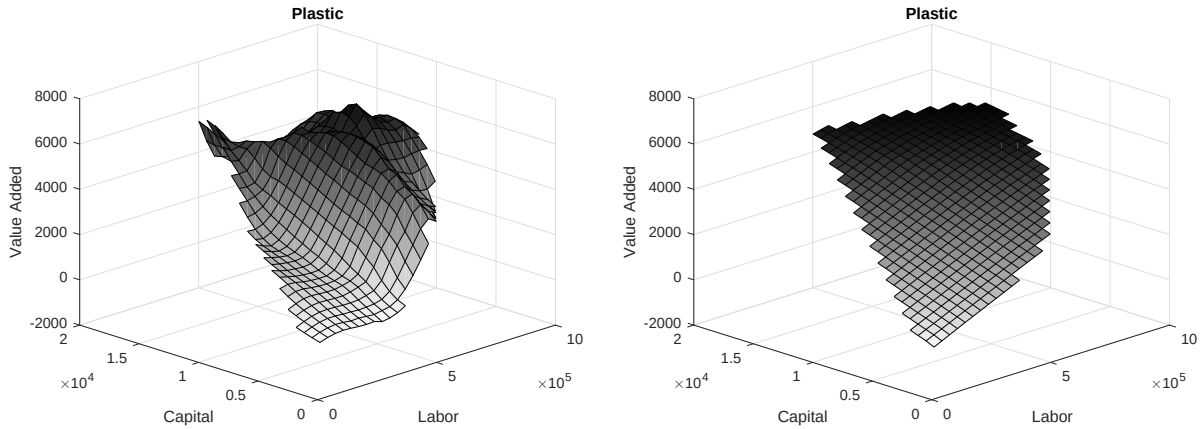
Table 7. SCKLS fitting statistics for cross sectional data.

Industry	Number of observations	R^2
Plastic	245	71.1%
Wood	132	43.8%

Table 8 reports additional information characterizing the production function: the marginal productivity and the marginal rates of substitution at the 10, 25, 50, 75 and 90 percentiles are reported for both measures. Here, the rate of substitution indicates how much labor is required to maintain the same level of output when we decrease a unit of capital. When comparing the two industries, we find that the wood industry has a larger marginal rate of substitution than the plastic industry. This indicates that capital is more critical in the wood industry than the plastic industry.

We also compare the estimated production function by the local linear and the SCKLS estimators. Figure 3 and Figure 4 show the estimated production function within the convex hull of observations for plastic and wood industries, respectively. Visually, the production

[7]We apply a Cobb–Douglas OLS to the second stage data $\{\boldsymbol{X}_j, y_j - \boldsymbol{Z}_j\boldsymbol{\gamma}\}_{j=1}^n$ which removes the effect of contextual variables from observed output. See Appendix F for details.



(a) Local Linear (b) SCKLS

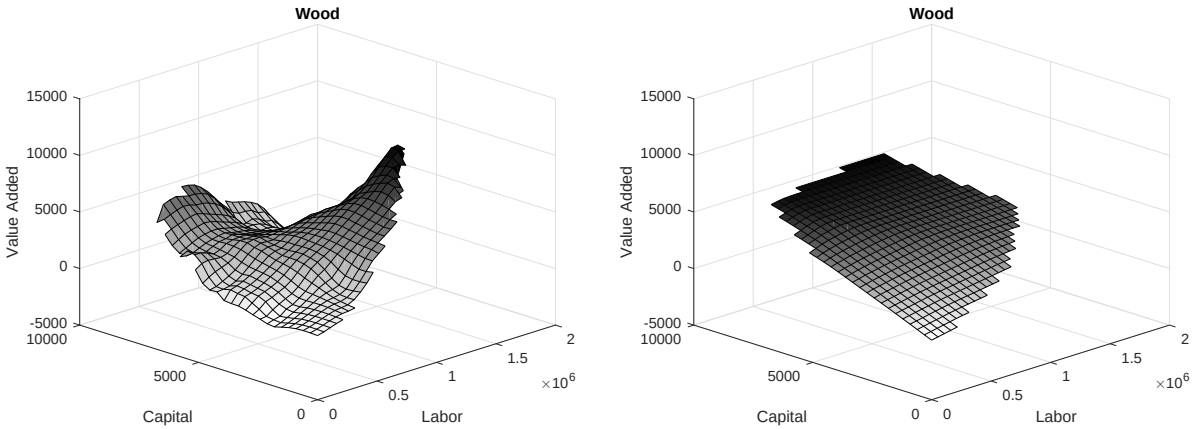
Figure 3. Production function estimated by LL and SCKLS for the plastic industry (2520)

function estimated by the LL estimator is difficult to interpret and the values of important economic quantities such as marginal products and marginal rates of substitution are also hard to interpret. In particular, it is not possible to identify most productive scale size.

Table 8. Characteristics of the production function.

	Plastic (2520)		
	Marginal Productivity		Marginal Rate of Substitution
	Labor ($= b_l$) (million peso/man hours)	Capital ($= b_k$) (peso/peso)	($= b_k/b_l$)
10th percentile	0.00396	0.111	23.3
25th percentile	0.00523	0.139	23.9
50th percentile	0.00579	0.139	24.0
75th percentile	0.00579	0.139	35.3
90th percentile	0.00579	0.260	44.8
	Wood (2010)		
	Marginal Productivity		Marginal Rate of Substitution
	Labor ($= b_l$)	Capital ($= b_k$)	($= b_k/b_l$)
10th percentile	1.46×10^{-18}	0.816	760
25th percentile	8.55×10^{-16}	0.816	760
50th percentile	0.00133	1.01	760
75th percentile	0.00133	1.01	9.73×10^{14}
90th percentile	0.00133	1.01	5.59×10^{17}

Table 9 reports the estimated coefficients for the exporting variables. In the plastic industry, the dummy variable for exporting is significant and positive while exports' share of sales is not. This indicates that the plants that export tend to produce more output than plants that do not export regardless of the export quantity. In contrast, the coefficient on the exports' share of sales is significant and positive in the wood industry while the dummy



(a) Local Linear (b) SCKLS

Figure 4. Production function estimated by LL and SCKLS shape constraints for the wood industry (2010)

variable for exporting is not significant, indicating that establishments in the wood industry tend to be more productive the more they export. Thus, in both industries we find evidence of increased productivity for exporting firms.

Table 9. Coefficient of contextual variables from a 2-stage model.

	Plastic (2520)		Wood (2010)	
	Dummy of exporting	Share of exporting in sales	Dummy of exporting	Share of exporting in sales
Point estimate	334.5	303.7	-763.0	4114
95% lower bound	148.7	-334.3	-1944	2568
95% upper bound	520.3	941.8	417.7	5660
p-value	4.70×10^{-4}	0.3493	0.2033	5.64×10^{-7}

Table 10 reports the most productive scale size for the 10, 25, 50, 75, 90 percentiles of Capital/Labor ratio distribution of observed input. In both industries, the observed value added output is the largest for establishments with high capital to labor ratios, indicating that capital-intensive establishments have increased actual output. Furthermore, labor-intensive establishments have smaller most productive scale size in both industries. This is consistent with the theory of the firm, i.e. firms grow and become more capital intensive over time by automating processes with capital and using less labor.

Table 10. Most productive scale size for each capital/labor ratio.

	Plastic (2520)		
Capital/Labor percentile	MPSS Labor	MPSS Capital	Output (Value added)
10th percentile	619580	519.1	3290
25th percentile	529980	1344	3010
50th percentile	529980	2604	3185
75th percentile	529980	5617	3602
90th percentile	529980	10270	4248
	Wood (2010)		
Capital/Labor percentile	MPSS Labor	MPSS Capital	Output (Value added)
10th percentile	2531100	741.6	1659
25th percentile	1045000	1200	2142
50th percentile	867250	2712	3470
75th percentile	662700	4179	4682
90th percentile	458150	5644	5893

7 Conclusion

This paper proposed the SCKLS estimator that imposes shape constraints on a local polynomial estimator. We show the consistency and convergence rate of this new estimator under monotonicity and concavity constraints, as well as its relationship with CNLS and CWB. We also illustrate how to use SCKLS to validate the imposed shape constraints. In applications where out-of-sample performance is less critical and the boundary behavior is of less concern, such as regulation applications, the CNLS estimator may be preferable because of its simplicity. In contrast, in cases where out-of-sample performance is important, such as survey data, the SCKLS estimator appears to be more robust. Simulation results reveal the SCKLS estimator outperforms CNLS and LL in most scenarios. We propose and validate the usefulness of several extensions, including variable bandwidth and non-uniform griding, which are important to estimate functions with non-uniform input data set which is common in manufacturing survey and census data. We also propose a test for the imposed shape constraints based on SCKLS. Finally, we demonstrate the SCKLS estimator empirically using Chilean manufacturing data. We compute marginal productivity, marginal rate of substitution, most productive scale size and the effects of exporting, and provide several economic insights.

One limitation of the proposed SCKLS estimator is its computation efficiency due to the

large number of constraints. The algorithm we proposed for reducing constraints performs well, and we demonstrate the ability to solve large problems instances within a reasonable time. Furthermore, our simulation results show good functional estimates even with a rough grid. Consequently, we can make use of the flexibility of the evaluation points to reduce the computational time of the estimator.

Potential future research could focus on the bandwidth selection methods. Typically, optimal bandwidth selection methods without shape constraints try to trade bias and variance to find the best estimator in terms of RMSE. Since the imposed shape restrictions already constrain the variance of the estimator to some extent, we expect that the optimal bandwidth in the SCKLS estimator will be smaller than the optimal unconstrained estimator. Further, if systematic inefficiency is present in the data, deconvoluting the residuals following the stochastic frontier literature would allow the investigation of a production frontier.

SUPPLEMENTARY MATERIAL

Appendix: The document contains: (A) extensions and the relationship between estimators; (B) technical proofs of the theoretical results; (C) a test of affinity using SCKLS; (D) an algorithm for SCKLS computational performance; (E) comprehensive results of existing and additional numerical experiments; (F) semiparametric model to integrate contextual variable; and (G) details of the application to the Chilean manufacturing data.

Appendix

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 $\{a_i, \boldsymbol{b}_i\}_{i=1}^m$. We re-define the SCKLS estimator as

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

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

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).

$$\boldsymbol{r} = (r_1, \ldots, r_d), \qquad \boldsymbol{r}! = r_1! \times \cdots r_d!, \qquad \bar{\boldsymbol{r}} = \sum_{k=1}^{d} r_k,$$

$$\boldsymbol{x}^{\boldsymbol{r}} = x_1^{r_1} \times \cdots x_d^{r_d}, \qquad \sum_{0 \le \bar{\boldsymbol{r}} \le p} = \sum_{k=0}^{p} \sum_{r_1=0}^{k} \cdots \sum_{r_d=0}^{k}, \qquad \text{and}$$

$$\left(D^{\boldsymbol{r}} g\right)(\boldsymbol{x}) = \frac{\partial^{\boldsymbol{r}} g(\boldsymbol{x})}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}$$

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

$$g(\boldsymbol{z}) := \sum_{0 \le \bar{\boldsymbol{r}} \le p} \frac{1}{\boldsymbol{r}!} \left(D^{\bar{\boldsymbol{r}}} g\right)(\boldsymbol{x}) (\boldsymbol{z} - \boldsymbol{x})^{\bar{\boldsymbol{r}}}. \tag{A.2}$$

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

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

where $\boldsymbol{b}_i$ is the functional or derivative estimates at each evaluation points and $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_m')'$. 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 $p \ge 3$ 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 burden[8]. Our experience suggests that

[8]While 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 establishments[9]. 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.

$$\begin{aligned} \min_{a_i, \boldsymbol{b_i}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}(y_j - a_i - (\boldsymbol{X_j} - \boldsymbol{x_i})'\boldsymbol{b_i})^2 w\left(\frac{\|\boldsymbol{X_j} - \boldsymbol{x_i}\|}{R_{\boldsymbol{x_i}}}\right) & \\ \text{subject to} \quad & a_i - a_l \geq \boldsymbol{b_i'}(\boldsymbol{x_i} - \boldsymbol{x_l}), & i, l = 1, \ldots, m \\ & \boldsymbol{b_i} \geq 0, & i = 1, \ldots, m \end{aligned} \tag{A.4}$$

where $w(\cdot)$ is a general weight function, $\|\cdot\|$ is the Euclidean norm and $R_{\boldsymbol{x_i}}$ denotes the Euclidean distance between $\boldsymbol{x_i}$ and k-th nearest neighbor of $\boldsymbol{x}_i$ among the set of all covariates $\{\boldsymbol{X}_j\}_{j=1}^n$. 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.

[9] An 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 $\{\boldsymbol{X}_j\}_{j=1}^n$.

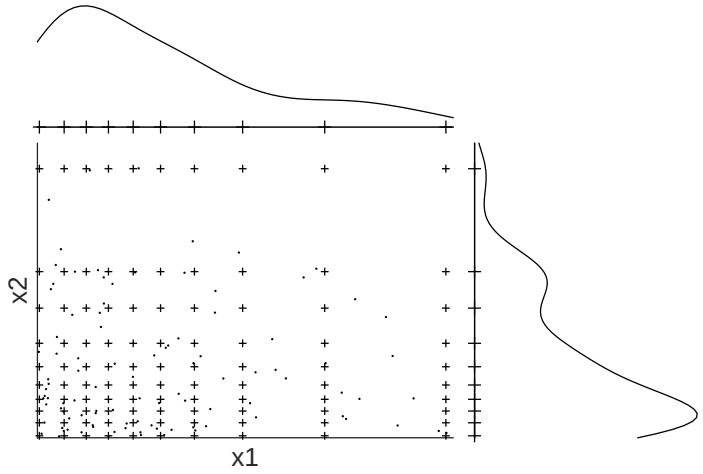


Figure A.1. Example of non-uniform grid with kernel density estimation.

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)[10]. 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 $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$, a set of unique fitted values, $\hat{y}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_{\boldsymbol{j}}\boldsymbol{X}_j$, can be found by solving the quadratic programming (QP) problem

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

where α_j and $\boldsymbol{\beta}_j$ 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

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

where $\mathcal{G}^n = \{g : \mathbb{R}^d \to \mathbb{R} \mid g(\boldsymbol{x}) = \min_{j\in\{1\ldots,n\}}(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}),\ \text{with } \boldsymbol{\beta}_j \geq 0 \text{ for } j = 1,\ldots,n\}$.

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

However, since the sets $\mathcal{G}^1, \mathcal{G}^2, \ldots$ 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

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

where $\boldsymbol{p} = (p_1, \ldots, p_n)'$, p_j is the weights introduced for each observation and $A_j(\boldsymbol{x})$ 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 $\boldsymbol{p}$ by minimizing its distance to unrestricted weights, $\boldsymbol{p}_u = (1/n, \ldots, 1/n)'$, under derivative-based shape constraints[11]. The problem is formulated as follows.

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

where $\boldsymbol{x_i}$ represents a set of points for evaluating constraints, the elements of $\boldsymbol{s}$ represent the order of partial derivative, and $g^{\boldsymbol{s}}(\boldsymbol{x}) = [\partial^{s_1} g(\boldsymbol{x}) \cdots \partial^{s_r} g(\boldsymbol{x})]/[\partial x_1^{s_1} \cdots \partial x_r^{s_r}]$ for $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)$. Here the shape restrictions (e.g. concavity/convexity and monotonicity constraints) are imposed at a set of evaluation points $\{\boldsymbol{x}_i\}_{i=1}^m$ through setting appropriate lower and upper bounds to the corresponding partial derivatives of the function. One way to interpret the CWB estimator is as a two-step process: 1) estimate an unconstrained kernel

[11] The 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.

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 $\boldsymbol{y}$ 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:

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

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

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

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

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 $\hat{g}^{(1)}(\boldsymbol{x}|\boldsymbol{p})$.

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_js are held fix. The 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

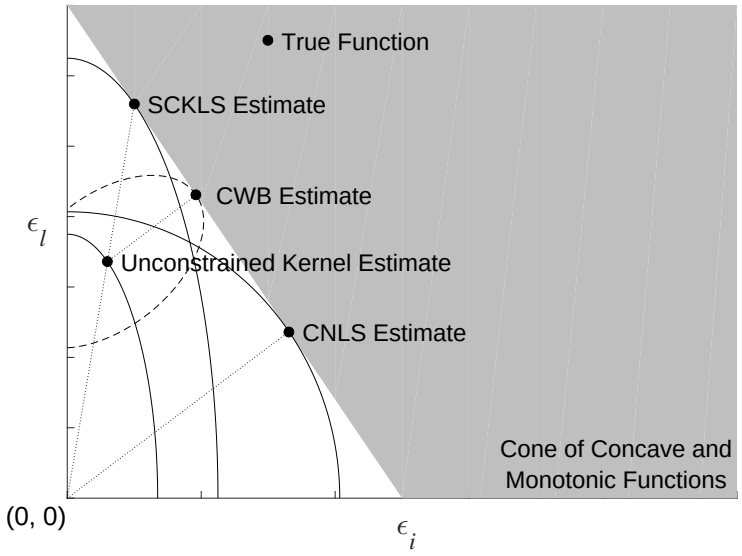


Figure A.2. Comparison of different estimators in the estimated-ϵ-space.

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* $\boldsymbol{x}_i = \boldsymbol{X}_i$ *for* $i = 1, \ldots, n$.

Proposition A.1. Suppose that Assumption A.1 holds. Then, for any n, when the vector of bandwidth goes to zero, i.e. $\|\boldsymbol{h}\| \to \mathbf{0}$ (where $\boldsymbol{h} = (h_1, \ldots, h_d)'$), the SCKLS estimator $\hat{g}_n$ converges to the CNLS estimator $\hat{g}_n^{CNLS}$ pointwise at $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

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 $\sup_{\boldsymbol{S}} \left|\hat{g}_n(\boldsymbol{x}) - g_0(\boldsymbol{x})\right| = o_p(1)$, while on the other hand, $\sup_{\boldsymbol{S}} \left|\hat{g}_n^{CNLS}(\boldsymbol{x}) - g_0(\boldsymbol{x})\right|$ 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. $\min_{k=1,\ldots,d} h_k \to \infty$), 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 $\hat{g}_n^{CWBY}$ 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. $\|\boldsymbol{h}\| \to \boldsymbol{0}$ (where $\boldsymbol{h} = (h_1, \ldots, h_d)'$), the SCKLS estimator $\hat{g}_n$ converges to the CWB in y-space estimator $\hat{g}_n^{CWBY}$ pointwise at $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

Proposition A.3 states that SCKLS and CWB in y-space estimators converge to the same estimates as $\|\boldsymbol{h}\| \to \boldsymbol{0}$. 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 $\boldsymbol{h}$, we have

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

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

$$\underset{a_1,\boldsymbol{b}_1,\ldots,a_n,\boldsymbol{b}_n}{\operatorname{argmin}} \sum_{j=1}^{n}(y_j - a_j)^2 K(\mathbf{0}) = \underset{a_1,\ldots,a_n}{\operatorname{argmin}} \sum_{j=1}^{n}(y_j - a_j)^2 = \underset{a_1,\ldots,a_n}{\operatorname{argmin}} L(g(a_j))$$

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 $D(\boldsymbol{p}) = \sum_{j=1}^{n}(p_u - p_j)^2 = \sum_{j=1}^{n}(1/n - p_j)^2 = L(m(g(p_j)))$. 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. $\|\boldsymbol{h}\| \to \mathbf{0}$ (where $\boldsymbol{h} = (h_1, \ldots, h_d)'$), 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 $\boldsymbol{X}$. The assumption on the rate of growth of m makes sure that the first partial derivative of SCKLS, $\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x})$, 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 $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ and evaluation points $\{\boldsymbol{x}_i\}_{i=1}^m$, the (unconstrained) local linear estimator at $\boldsymbol{x}_i$ is $(\tilde{a}_i, \tilde{\boldsymbol{b}}_i)$ for $i = 1, \ldots, m$, where $(\tilde{a}_1, \tilde{\boldsymbol{b}}_1, \ldots, \tilde{a}_m, \tilde{\boldsymbol{b}}_m)$

is the (unique) minimizer of

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

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

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

where

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

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

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

Therefore, SCKLS can be simply viewed as a minimizer of

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

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

$$\|g_1 - g_2\|_{n,m}^2 = \frac{1}{m}\sum_{i=1}^{m} \Big(g_1(\boldsymbol{x}_i) - g_2(\boldsymbol{x}_i), \big(\frac{\partial g_1}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \frac{\partial g_2}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)'h\Big)\boldsymbol{\Sigma}_i \begin{pmatrix} g_1(\boldsymbol{x}_i) - g_2(\boldsymbol{x}_i) \\ \big(\frac{\partial g_1}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \frac{\partial g_2}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)'h \end{pmatrix},$$

then SCKLS belongs to[12]

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

[12]To 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.

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

Below, we list some useful results on the behaviors of $\boldsymbol{\Sigma}_i$ and $(\tilde{a}_i, \tilde{\boldsymbol{b}}_i)$. 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 $\boldsymbol{\Sigma}_i$ are in $[1/C, C]$ for all $i = 1, \ldots, m$ 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 $n \to \infty$,*

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

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 $\hat{g}_n(\boldsymbol{x}_i)$ and $\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)$ for $i = 1, \ldots, m$ 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

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

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,

$$\|\hat{g}_n - g_0\|_{n,m}^2 \leq 4\|\tilde{g}_n - g_0\|_{n,m}^2. \tag{A.9}$$

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

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

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

$$\begin{aligned} \|\hat{g}_n - g_0\|_{n,m}^2 &= \frac{1}{m}\sum_{i=1}^{m} \Big(\hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i), \big(\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \frac{\partial g_0}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)' h\Big) \boldsymbol{\Sigma}_i \begin{pmatrix} \hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i) \\ \big(\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \frac{\partial g_0}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big) h \end{pmatrix} \\ &\geq \frac{1}{Cm}\sum_{i=1}^{m} (\hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i))^2, \end{aligned} \tag{A.10}$$

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

Plugging the above two equations into (A.9) yields

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

□

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

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

implies that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

violating

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

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

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

implies that

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

as $n \to \infty$.

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

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

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

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 $\boldsymbol{C} \subset \text{int}(\boldsymbol{S})$ (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 $\boldsymbol{C}$ with probability one as $n \to \infty$.*

Proof. As explained before, here we focus on the scenario of $d = 1$. Without loss of generality, we can take $\boldsymbol{S} = [0, 1]$ and $\boldsymbol{C} = [\delta, 1 - \delta]$ 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 $n \to \infty$.

Since $\hat{g}_n$ is increasing, $\sup_{x \in [\delta, 1-\delta]} |\hat{g}_n(x)| = \max\Big(|\hat{g}_n(\delta)|, |\hat{g}_n(1-\delta)|\Big)$. 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

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

where $q(\cdot)$ is the density function with respect to what the empirical distribution of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ converges to (see Assumption 2(i)). Here the last line also follows from Assumption 2(i). Note that Theorem 1 says that $\frac{1}{m}\sum_{i=1}^{m}(\hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i))^2 = o_p(1)$, 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 $n \to \infty$.

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 $n \to \infty$.

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

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

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). □

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 $\boldsymbol{C} \subset \text{int}(\boldsymbol{C'})$ and $\boldsymbol{C'} \subset \text{int}(\boldsymbol{S})$, 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 $\boldsymbol{C'}$ with probability one as $n \to \infty$. Therefore, $\{\hat{g}_n(\boldsymbol{x}) - g_0(\boldsymbol{x})\}^2 \mathbf{1}_{\{\boldsymbol{x}\in\boldsymbol{C'}\}}$ belongs to the class of functions that is bounded and equicontinuous over $\boldsymbol{C'}$. 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

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

In addition, it follows from Theorem 1 that

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

Combining the above two equations together yields

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

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

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

Second claim: when m is fixed.

In views of Lemma B.1 and Theorem 1,

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

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 $|\hat{g}_n(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)| = O_p(n^{-2/(4+d)}\log n) \xrightarrow{p} 0$ and $\|\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \frac{\partial g_0}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\| = O_p(n^{-1/(4+d)}\log n) \xrightarrow{p} 0$ for every $i = 1, \ldots, m$. □

B.3.3 Proof of Theorem 3

Proof. Using Equation (A.10) but focusing on the difference between the derivatives instead, we have that

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

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

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

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

$$\hat{g}_n(\boldsymbol{x}) = \min_{i\in\{1,\ldots,m\}} \Big\{ \hat{g}_n(\boldsymbol{x}_i) + (\boldsymbol{x} - \boldsymbol{x_i})' \frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) \Big\},$$

we have that with probability one,

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

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

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

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

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

□

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

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

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

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

$$
\frac{2}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i\begin{pmatrix}\hat{a}_i - g_0^*(\boldsymbol{x}_i)\\ (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h\end{pmatrix} \tag{A.11}
$$

$$
\geq \frac{1}{m}\sum_{i=1}^{m}\big(g_0^*(\boldsymbol{x}_i) - \hat{a}_i, (\frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \hat{\boldsymbol{b}}_i)'h\big)\boldsymbol{\Sigma}_i\begin{pmatrix}g_0^*(\boldsymbol{x}_i) - \hat{a}_i\\ (\frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \hat{\boldsymbol{b}}_i)h\end{pmatrix} \geq \frac{1}{mC}\sum_{i=1}^{m}(g_0^*(\boldsymbol{x_i}) - \hat{a}_i)^2 \tag{A.12}
$$

Next, we show that the quantity in (A.11) converges to zero in probability as $n \to \infty$. The proof can be divided into six steps:

1. The contribution to (A.11) from evaluation points lying outside a carefully pre-chosen compact subset $\boldsymbol{S}'$ of the interior of $\boldsymbol{S}$ (denoted as $\text{int}(\boldsymbol{S})$) can be made arbitrarily small.

This follows from the Cauchy–Schwarz inequality that

$$\frac{1}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix} \mathbf{1}_{\{\boldsymbol{x}\notin \boldsymbol{S}'\}}$$

$$\leq \sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \tilde{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix} \mathbf{1}_{\{\boldsymbol{x}\notin \boldsymbol{S}'\}}} \tag{A.13}$$

$$\times \sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\hat{a}_i - g_0^*(\boldsymbol{x}_i), (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix}}. \tag{A.14}$$

Because of Lemma B.1 and Assumption 2(i), the quantity in (A.13) can be made arbitrarily small by choosing $\boldsymbol{S}'$ sufficiently close to $\boldsymbol{S}$. In addition, applying the Cauchy–Schwarz inequality to (A.11) and comparing it to (A.12) yields

$$2\sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \tilde{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix}}$$

$$\times \sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\hat{a}_i - g_0^*(\boldsymbol{x}_i), (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix}}$$

$$\geq \frac{1}{m}\sum_{i=1}^{m}\big(g_0^*(\boldsymbol{x}_i) - \hat{a}_i, (\frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \hat{\boldsymbol{b}}_i)'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} g_0^*(\boldsymbol{x}_i) - \hat{a}_i \\ (\frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i) - \hat{\boldsymbol{b}}_i)h \end{pmatrix},$$

so (A.14) is no greater than

$$2\sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \tilde{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix}}$$

$$\to 2\Big\{\int_{\boldsymbol{S}}(g_0(\boldsymbol{x}) - g_0^*(\boldsymbol{x}))^2 Q(d\boldsymbol{x})\Big\}^{1/2} \leq 2\Big\{\int_{\boldsymbol{S}} g_0^2(\boldsymbol{x}) Q(d\boldsymbol{x})\Big\}^{1/2}.$$

Consequently, the claim in this step is proved.

2. We now investigate the contribution to (A.11) from evaluation points lying inside $\boldsymbol{S}'$. Using Lemma B.5, we have that $\hat{g}_n$ is bounded (i.e. from both below and above) and M-Lipschitz over $\boldsymbol{S}'$ in probability.

Combining this with Lemma B.1 implies that

$$\left| \frac{1}{m}\sum_{i=1}^{m} \big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ \big(\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)h \end{pmatrix} \mathbf{1}_{\{\boldsymbol{x}\in \boldsymbol{S}'\}} \right.$$
$$\left. - \frac{1}{m}\sum_{i=1}^{m} \left((g_0 - g_0^*)(\boldsymbol{x}_i), (\frac{\partial (g_0 - g_0^*)}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\right)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ \big(\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)h \end{pmatrix} \mathbf{1}_{\{\boldsymbol{x}\in \boldsymbol{S}'\}} \right| \to 0$$

in probability. As such, we can instead work on

$$\frac{1}{m}\sum_{i=1}^{m} \left((g_0 - g_0^*)(\boldsymbol{x}_i), (\frac{\partial (g_0 - g_0^*)}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\right)\boldsymbol{\Sigma}_i \begin{pmatrix} \hat{a}_i - g_0^*(\boldsymbol{x}_i) \\ \big(\hat{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\big)h \end{pmatrix} \mathbf{1}_{\{\boldsymbol{x}\in \boldsymbol{S}'\}} \tag{A.15}$$

3. Next, we bound (and eliminate) the influence from the parts involving partial derivatives of g_0, g_0^* and $\hat{g}_n$ in (A.15). Since $\hat{g}_n$ is bounded and M-Lipschitz over $\boldsymbol{S'}$ in probability, together with Lemma B.2, we could bound (A.15) from above by

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

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

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

$$\hat{g}_n^{\boldsymbol{S}'}(\boldsymbol{x}) = \min_{i\in\{1,\ldots,m|\boldsymbol{x}_i\in \boldsymbol{S}'\},} \Big\{\hat{g}_n(\boldsymbol{x}_i) + (\boldsymbol{x} - \boldsymbol{x_i})'\frac{\partial \hat{g}_n}{\partial \boldsymbol{x}}(\boldsymbol{x}_i)\Big\}.$$

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

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

over $\boldsymbol{C}$ in probability via Lemma B.5. Then it follows that

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

in probability as $n \to \infty$.

- The function $\{(g_0 - g_0^*)(\hat{g}_n - g_0^*)\}(\cdot)$ is bounded and Lipschitz over $\boldsymbol{S}'$ in probability (where the constants do not depend on n). So is $\{(g_0 - g_0^*)(\hat{g}_n^{\boldsymbol{S}'} - g_0^*)\}(\cdot)$ over $\boldsymbol{S}$. This also means that $\{(g_0 - g_0^*)(\hat{g}_n^{\boldsymbol{S}'} - g_0^*)\}(\cdot)$ is equicontinuous over $\boldsymbol{S}$.

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

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

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 $n \to \infty$) is at most

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

Note that g_0^* minimizes

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

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 $\epsilon \to 0$) for all $g \in G_2$.

- Both |(II)| and |(III)| in the limit can be arbitrarily small for $\boldsymbol{S}'$ sufficiently close to $\boldsymbol{S}$. 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 $\boldsymbol{S}'$ such that the quantity in (A.11) is no bigger than ϵ in probability as $n \to \infty$. Since the quantity in (A.11) is also non-negative, our claim that (A.11) converges to zero in probability is verified.

Finally, uniform consistency over any $\boldsymbol{C}$ 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. □

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 $\hat{g}_n^\circ$ and $\tilde{g}_n^\circ$ the SCKLS and LL estimators based on the same covariates, evaluation points and bandwidth used in calculating T_n, but with the response vector $(\epsilon_1, \ldots, \epsilon_n)'$ (instead of $\boldsymbol{y}_n$) 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 $k = 1, \ldots, B$, $T_{nk} = \|\hat{g}_{nk} - \tilde{g}_{nk}\|_{n,m}$, 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 $(u_{1k}\tilde{\epsilon}_1, \ldots, u_{nk}\tilde{\epsilon}_n)'$. 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 $\hat{g}_{nk}^\circ$ and $\tilde{g}_{nk}^\circ$ are the SCKLS and LL estimators based on the same covariates, evaluation points and bandwidth used in calculating T_n, but with the response $(u_{1k}\epsilon_1, \ldots, u_{nk}\epsilon_n)'$. Let $\boldsymbol{e} = (|\epsilon_1|, \ldots, |\epsilon_n|)'$ and denote $p_n^\circ = \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}_{\{T_n^\circ \leq T_{nk}^\circ\}}$. Then, it follows from the symmetry of the error distribution that conditioning on the values of the absolute errors

(i.e. $(|\epsilon_1|, \ldots, |\epsilon_n|)' = \boldsymbol{e}$), the quantities

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

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

$$P(p_n^\circ \leq \alpha) = E\Big\{ P\Big(p_n^\circ \leq \alpha \Big| (|\epsilon_1|, \ldots, |\epsilon_n|)' = \boldsymbol{e}\Big)\Big\} \leq \frac{\lfloor B\alpha \rfloor + 1}{1+B} \to \alpha.$$

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,

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

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 $\hat{g}_n^\circ$ and $\tilde{g}_n^\circ$ 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 $(g_0(\boldsymbol{X}_1), \ldots, g_0(\boldsymbol{X}_n))'$. 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 $k = 1, \ldots, B$,

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

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

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

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

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

  Note that

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

  Here we have that $\tilde{g}_n \to g_0$ by Fan and Guerre (2016) and $\hat{g}_n \to g_0^*$ over $\boldsymbol{S}'$ by our Theorem 4. Since $\tilde{g}_n - \hat{g}_n$ is Lipschitz over $\boldsymbol{S}'$, 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 $\delta^2 Q(\boldsymbol{S}')$ in the limit as $n \to \infty$ (also $h \to 0$). 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 $\boldsymbol{S}$ (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 $W_{n1}, \ldots, W_{nB}$ are exchangeable. Thus, for any $\alpha \in (0, 1)$, as $n \to \infty$,

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

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 $\boldsymbol{h}$, we have

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

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

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

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

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

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

B.5.2 Proof of Proposition A.2

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

$$K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right) = K(\boldsymbol{0}) \quad \text{for } \forall i, j. \tag{A.16}$$

By substituting (A.16) into the objective function of (3) converges to

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

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

$$\underset{a_1,\boldsymbol{b}_1,\ldots,a_m,\boldsymbol{b}_m}{\operatorname{argmin}} \sum_{i=1}^{m}\sum_{j=1}^{n}(y_j - a_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)'\boldsymbol{b}_i)^2 \tag{A.17}$$

subject to constraints. Rewrite $a_i + (\boldsymbol{X}_j - \boldsymbol{x}_i)'\boldsymbol{b_i} = \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{X}_j$ for $i = 1,\ldots,m$ and $j = 1,\ldots,n$. Then the objective function of (3) can be rewritten as follows with (A.17).

$$\begin{aligned}
\min_{\alpha_1,\boldsymbol{\beta}_1,\ldots,\alpha_m,\boldsymbol{\beta}_m} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n}(y_j - (\alpha_i + \boldsymbol{\beta}_i'\boldsymbol{X}_j))^2 \\
\text{subject to} \quad & \alpha_i + \boldsymbol{\beta}_i'\boldsymbol{x}_i \le \alpha_l + \boldsymbol{\beta}_l'\boldsymbol{x}_i \qquad i,l = 1,\ldots,m \\
& \boldsymbol{\beta}_i \ge 0 \qquad\qquad\qquad\qquad\quad i = 1,\ldots,m
\end{aligned}$$

Here, since we do not impose any weight on the objective function, it is easy to see that $\alpha_1 = \cdots = \alpha_m$ and $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m$. Then the Afriat constraints become redundant, resulting in

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

□

B.5.3 Proof of Proposition A.3

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

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

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

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

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

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

and thus

$$\hat{g}(\boldsymbol{x}_i|\boldsymbol{p}) = \sum_{j=1}^{n} p_j A_j(\boldsymbol{X}_i) y_j = n p_i y_i \quad \forall i = 1, \ldots, n. \tag{A.18}$$

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

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & D_y(\boldsymbol{p}) = \sum_{i=1}^{n} (y_i - n p_i y_i)^2 \\ \text{subject to} \quad & l(\boldsymbol{x}_i) \leq \hat{g}^{(\boldsymbol{s})}(\boldsymbol{x}_i|\boldsymbol{p}) \leq u(\boldsymbol{x}_i), \quad i = 1, \ldots, n. \end{aligned} \tag{A.19}$$

Recognize that if $\hat{g}_n = n p_i y_i$ 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 $\hat{g}_n = n p_i y_i$ is simply a system of n equations and n unknowns. □

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

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

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 $\hat{g}_n^{\mathrm{V}}$, the SCKLS estimator with only a set of convexity

constraints as

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

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

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

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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ and bandwidth $\boldsymbol{h}_n$). The resulting estimators are denoted by $\hat{g}_n^{\mathrm{V}}$ and $\hat{g}_n^{\Lambda}$, where $\hat{g}_n^{\mathrm{V}}$ is the SCKLS estimator using only a set of convexity constraints, while $\hat{g}_n^{\Lambda}$ is the SCKLS estimator using only a set of concavity constraints, all based on $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$.We then define the test statistics to be

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

2. We simulate the distributional behavior of the test statistics B times under H_0. For $k = 1, \ldots, B$, we set the observations to be $\{\boldsymbol{X}_j, y_{jk}\}_{j=1}^n$ (i.e. no change in the values of the covariates), where $\boldsymbol{y}_{nk} = (y_{1k}, \ldots, y_{nk})'$ 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 $\boldsymbol{y}_{nk}$ 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 $\hat{g}_{nk}^{\mathrm{V}}$ and $\hat{g}_{nk}^{\Lambda}$, where $\hat{g}_{nk}^{\mathrm{V}}$ is the SCKLS estimator using only the convexity constraint, while $\hat{g}_{nk}^{\Lambda}$ is the SCKLS estimator using only the concavity

constraint, all based on $\{\boldsymbol{X}_j, y_{jk}\}_{j=1}^n$. So

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

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, $\hat{g}_n^{\mathrm{V}}$ and $\hat{g}_n^{\Lambda}$, 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 $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ and $\{\boldsymbol{X}_j, \epsilon_j\}_{j=1}^n$ 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 $\hat{g}_n^{\mathrm{V}}$ by the SCKLS with both the convexity and monotonicity constraints, and replace $\hat{g}_n^{\Lambda}$ 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:

$$g_0(\boldsymbol{x}) = \frac{1}{d}\sum_{k=1}^{d} x_k^p \tag{A.20}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)'$. With n observations, for each pair $(\boldsymbol{X}_j, y_j)$, each component of the input, $\boldsymbol{X}_{jk}$, is randomly and independently drawn from uniform distribution $unif[0,1]$, 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 A.4 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.

Table A.4. 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

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 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 A.3. Further, we summarize our implementation of the algorithm proposed in Lee et al. (2013) below and label it as Algorithm 1.

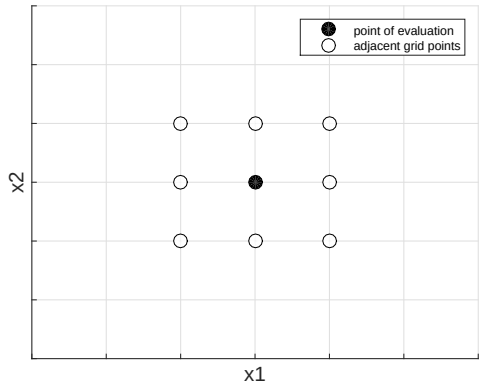


Figure A.3. Definition of adjacent grid in two-dimensional case.

Algorithm 1 Iterative approach for SCKLS computational speedup

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

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 CWB[13].

For CWB estimator, we use a local linear estimator to obtain the weighting matrix $A_j(\boldsymbol{x})$ in (A.6). The first partial derivative of $\hat{g}(\boldsymbol{x}|\boldsymbol{p})$ is obtained by approximating the derivatives through numerical differentiation $\hat{g}^{(1)}(\boldsymbol{x}|\boldsymbol{p}) = \frac{\hat{g}(\boldsymbol{x}+\Delta|\boldsymbol{p})-\hat{g}(\boldsymbol{x}|\boldsymbol{p})}{\Delta}$, where Δ is a small positive constant[14].

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 A.5 and Table A.6 show the RMSE of Experiment 1 on observation points and evaluation points respectively.

Table A.7 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 $\boldsymbol{h} \in [0, 5] \times [0, 5]$ 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

[13]For CWB, `SeDuMi` provides a better solution than `quadprog`, while both `SeDuMi` and `quadprog` give exactly the same solution for SCKLS.

[14]Du et al. (2013) proposes to use an analytical derivative for the first partial derivative of $\hat{g}(\boldsymbol{x}|\boldsymbol{p})$; 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.

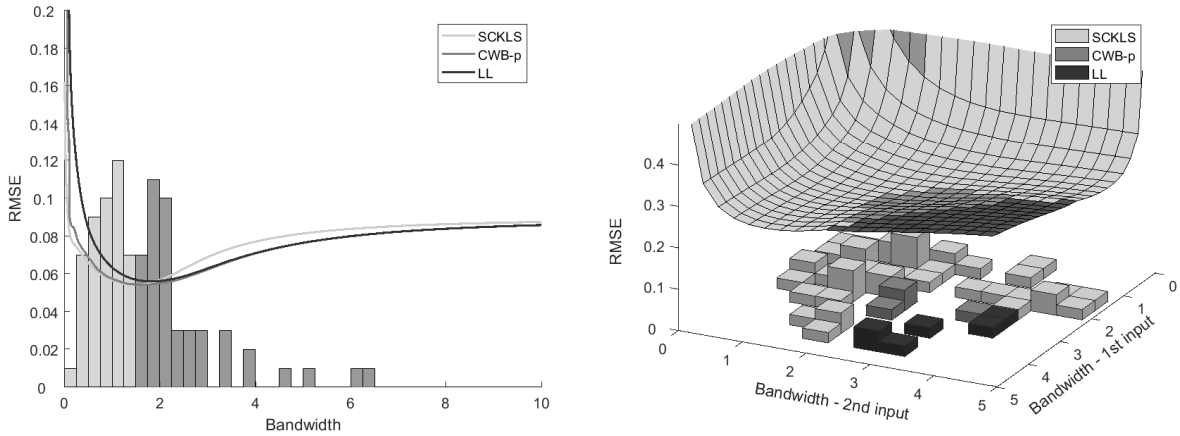
Table A.5. RMSE on observation points for Experiment 1

	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 A.6. 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 A.4. The histogram shows the distribution of bandwidths selected by LOOCV. The curves show the relative performance of each estimator.

Table A.7. Computational time for Experiment 1

		Average of computational time in seconds; (percentage of Afriat constraints included in the final optimization problem)				
Number of observations		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,

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

For each pair $(\boldsymbol{X}_j, y_j)$, each component of the input, $\boldsymbol{X}_{jk}$, 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 A.8 and Table A.9 show the RMSE of this experiment on observation points and evaluation points respectively.

Table A.8. 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 A.9. 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 A.10 and Table A.11 show the RMSEs of Experiment 2 on observation points and evaluation points respectively. In addition, Table A.12 shows the computational time of Experiment 2 for each estimator.

Table A.10. 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 A.11. 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 A.12. Computational time for Experiment 2

		Average of computational time in seconds; (percentage of Afriat constraints included in the final optimization)		
Number of evaluation points		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,

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

For each pair $(\boldsymbol{X}_j, y_j)$, each component of the input, $\boldsymbol{X}_{jk}$, 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 A.13 and Table A.14 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 A.13. 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 A.14. 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 $j = 1, \ldots, n$, input, $\boldsymbol{X}_j = (X_{j1}, X_{j2})'$, 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 A.15 and Table A.16 show the RMSEs of this experiment on observation points and evaluation points. Figure A.5 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 A.15. RMSE on observation points for Experiment: misspecified shape

	Average of RMSE on observation points				
Number of observations	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 A.16. 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

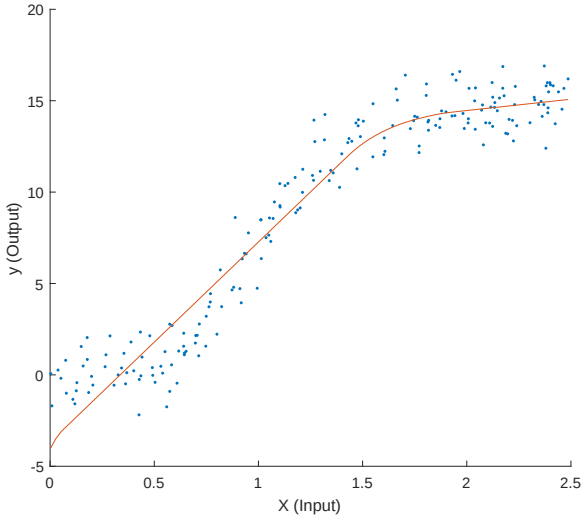


Figure A.5. A typical run of SCKLS when the truth is S-shaped.

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:

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

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

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

where $\tilde{\boldsymbol{Z}}_j = \boldsymbol{Z}_j - \hat{E}[\boldsymbol{Z}_j|\boldsymbol{X}_j]$ and $\tilde{y}_j = y_j - \hat{E}[y_j|\boldsymbol{X}_j]$ respectively, and each conditional expectation is estimated by kernel estimation method such as local linear. Finally, we apply the SCKLS estimator to the data $\{\boldsymbol{X}_j, y_j - \boldsymbol{Z}_j'\hat{\boldsymbol{\gamma}}\}_{j=1}^n$. 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 $\boldsymbol{Z}_j$ to the estimation performance by comparing SCKLS with and without contextual variables. We use two different Cobb–Douglas production functions as the true DGP:

$$g_0(\boldsymbol{x}, z) = \prod_{k=1}^{d} x_k^{\frac{0.8}{d}} + z\gamma, \tag{A.21}$$

$$g_0(\boldsymbol{x}) = \prod_{k=1}^{d} x_k^{\frac{0.8}{d}}, \tag{A.22}$$

where for each $(\boldsymbol{X}_j, Z_j, y_j)$, the contextual variable Z_j is a scalar value independent of $\boldsymbol{X}_j$ drawn randomly and independently from $uinf[0,1]$, 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 (A.21) and (A.22), respectively.

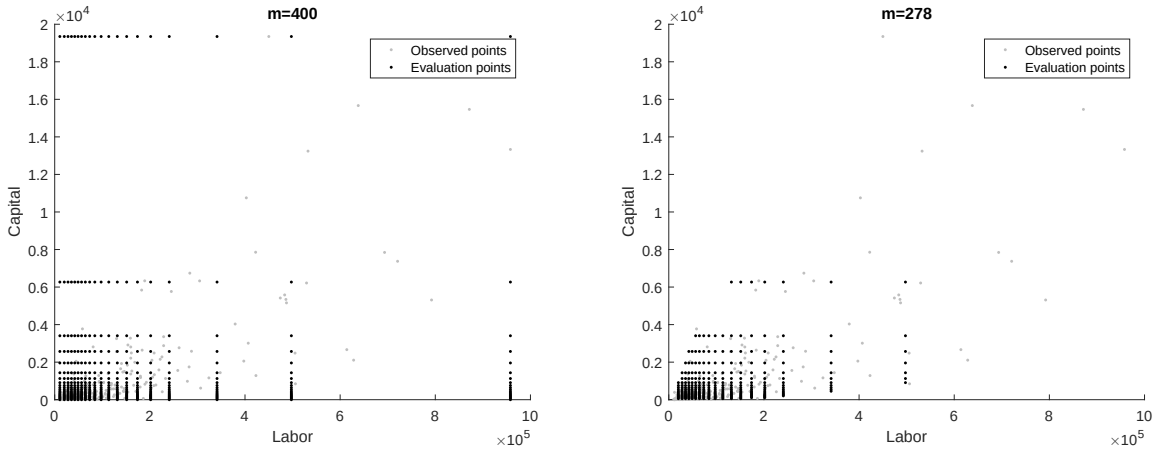
Table A.17 and Table A.18 show the RMSEs of this experiment on observation points and evaluation points respectively. The RMSE is obtained by comparing estimates of production function and the true production function. We see that having extra contextual variables does not deteriorate the performance 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 A.17. RMSE on observation points for experiments with/without Z-variable

		Average of RMSE on observation points				
Number of observations		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 A.18. RMSE on evaluation points for experiments with/without Z-variable

		Average of RMSE on evaluation points				
Number of observations		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



(a) Before deletion (b) After deletion

Figure A.6. Proposed evaluation points with Plastic industry (2520)

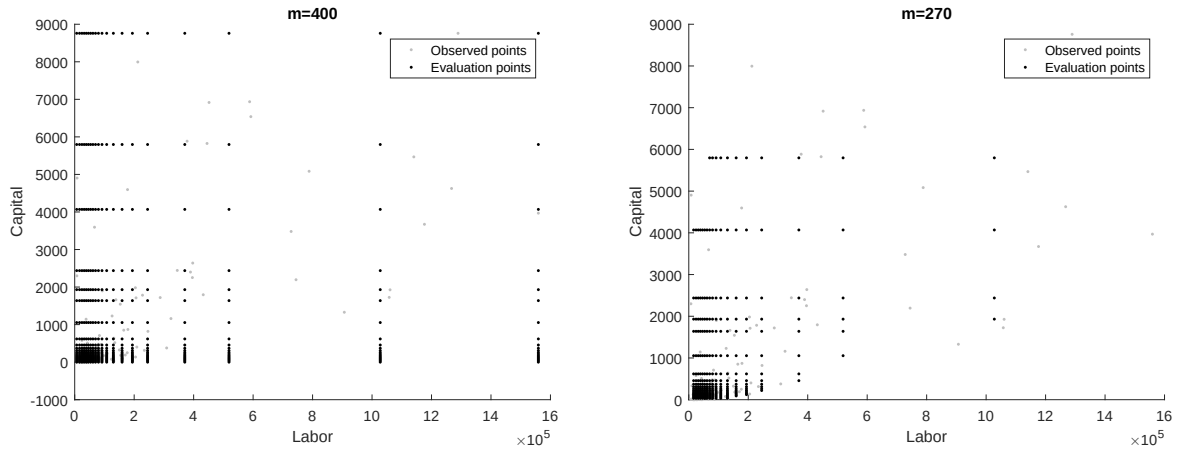
G Details on the application to the Chilean manufacturing data

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 $\{\boldsymbol{X}_j\}_{j=1}^n$. Figure A.6 and Figure A.7 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 $\{\boldsymbol{X}_j\}_{j=1}^n$, the number is $m \approx 270$ for both industries.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review 13*(3), 568–598.



(a) Before deletion (b) After deletion

Figure A.7. Proposed evaluation points with Wood industry (2010)

Alvarez, R. and H. Görg (2009). Multinationals and plant exit: Evidence from chile. *International Review of Economics & Finance 18*(1), 45–51.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica 68*(2), 399–405.

Banker, R. D. and A. Maindiratta (1992). Maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis 3*(4), 401–415.

Benavente, J. M. (2006). The role of research and innovation in promoting productivity in chile. *Economics of Innovation and New Technology 15*(4-5), 301–315.

Beresteanu, A. (2005). Nonparametric analysis of cost complementarities in the telecommunications industry. *RAND Journal of Economics 36*(4), 870–889.

Beresteanu, A. (2007). Nonparametric estimation of regression functions under restrictions on partial derivatives. Working paper.

Bernard, A. B. and J. B. Jensen (2004). Exporting and productivity in the usa. *Oxford Review of Economic Policy 20*(3), 343–357.

Bertsekas, D. (1995). *Nonlinear Programming*. Athena Scientific.

Birke, M. and H. Dette (2007). Estimating a convex function in nonparametric regression. *Scandinavian Journal of Statistics 34*(2), 384–404.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics 26*(4), 607–616.

Carroll, R. J., A. Delaigle, and P. Hall (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association 106*(493), 191–202.

Cavaliere, G., H. Bohn Nielsen, and A. Rahbek (2017). On the consistency of bootstrap testing for a parameter on the boundary of the parameter space. *Journal of Time Series Analysis 38*, 513534.

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.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association 74*(368), 829–836.

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.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*(1), 162 – 169.

De Loecker, J. (2007). Do exports generate higher productivity? evidence from slovenia. *Journal of International Economics 73*(1), 69–98.

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. (1956). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal 1956*(2), 125–153.

Grenander, U. (1981). *Abstract Inference*. John Wiley & Sons.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics 29*(6), 1653–1698.

Hall, P. and N. E. Heckman (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics 28*(1), 20–39.

Hall, P. and L.-S. Huang (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics 29*(3), 624–647.

Hanson, D. and G. Pledger (1976). Consistency in concave regression. *The Annals of Statistics 4*(6), 1038–1050.

Henderson, D. J. and C. F. Parmeter (2015). *Applied nonparametric econometrics*. Cambridge University Press.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association 49*(267), 598–619.

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 stonezd 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.

Kuosmanen, T. and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *Journal of Productivity Analysis 38*(1), 11–28.

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.

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies 70*(2), 317–341.

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.

Liu, R. Y. (1988, 12). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics 16*(4), 1696–1708.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *The Annals of Statistics 19*(2), 741–759.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics 21*(1), 255–285.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Jounral of Time Series Analysis 17*(6), 571–599.

Mazumder, R., A. Choudhury, G. Iyengar, and B. Sen (2015). A Computational Framework for Multivariate Convex Regression and its Variants. arXiv preprint arXiv:1509.08165.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming 103*(1), 127–152.

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.

Pavcnik, N. (2002). Trade liberalization, exit, and productivity improvements: Evidence from chilean plants. *The Review of Economic Studies 69*(1), 245–276.

Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics 119*(1), 99–130.

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.

Seijo, E. and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics 39*(3), 1633–1657.

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.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*(4), 595–620.

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.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*(4), 1261–1295.