THEORY AND APPLICATION OF LOCAL WEIGHTED SHAPE CONSTRAINED ESTIMATORS FOR ANALYZING CENSUS OF MANUFACTURING DATA

A Dissertation

by

DAISUKE YAGI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Andrew L. Johnson
Committee Members,	Yu Ding
	Erick Moreno-Centeno
	Jianhua Huang
Head of Department,	Mark Lawley

August 2018

Major Subject: Industrial Engineering

ABSTRACT

Efficiency and productivity analysis focuses on firm performance to obtain firm–level and industry–level economic structural insights. This study provides the theoretical and methodological basis for nonparametric production function estimation using local weighting and imposing shape constraints to avoid functional misspecification and to improve the interpretability of estimation results.

The first contribution is a model that combines a conventional local weighted estimator with monotonicity and global concavity constraints consistent with a production process with decreasing returns to scale. The second contribution is a model that imposes more complicated shape constraints allowing small firms to benefit from increasing returns to scale while still imposing decreasing returns to scale for large firms. This set of shape constraints is referred to as an S-shape production function and the relationship to the Regular Ultra Passum law is described. Further, an algorithm is proposed to estimate a production function satisfying the S-shape restriction, convex input sets and allowing for potentially non-homothetic input isoquants. The third contribution is a model that further extends the first two contributions to address the simultaneity issue using an instrumental variables approach. The proposed model imposes shape constraints in a Landweber–Fridman regularization. In addition to methodological contributions, simulation and application results are provided to demonstrate the improved finite sample performance and the interpretability of estimation results. Insights are gained for both Chilean and Japanese manufacturing by using the census of manufacturing data from these two countries.

DEDICATION

To all my family members, especially to my father, Tetsu Yagi; and mother, Yoko Yagi.

ACKNOWLEDGMENTS

I would first like to express my gratitude to my advisor, Dr. Andrew Johnson, for your continuous support. It would be impossible to overcome a lot of tough challenges without your advice. Whenever I discussed with you, your creativity inspired me to come up with some solutions for severe problems. I would also like to thank my committee members, Dr. Ding, Dr. Moreno and Dr. Jianhua for giving me insightful advice. I would appreciate co–authors of my paper, Dr. Chen and Dr. Kuosmanen, for giving me your help and advice.

I would also like to thank all my friends and colleagues who have helped and inspired me over five years.

Finally, I greatly appreciate and acknowledge the love and support I receive from my family.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Andrew L. Johnson, Dr. Yu Ding, and Dr. Erick Moreno-Centeno of the Department of Industrial and Systems Engineering and Dr. Jianhua Huang of the Department of Statistics.

The data analyzed for Chapter 3 and 4 was provided by Dr. Hiroshi Morita of the Department of Information and Physical Sciences at Osaka University. The statistical theory in Chapter 2 and 3 were supported by Dr. Yining Chen of the Department of Statistics at London School of Economics and Political Science. The concept and motivation of Chapter 2 was developed in collaboration with Dr. Timo Kuosmanen of the School of Business at Aalto University. And my advisor Dr. Andrew Johnson provided various support throughout all chapters.

All work on the dissertation is the result of the student's leadership, execution and collaboration with the above stated participates.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

An aspect of productivity analysis is the study of firm performance by comparing the outputs produced to the inputs consumed. Specifically, we are interested in estimating a production function, which describes the relationship between observed outputs, such as value added, and factors of production, such as labor and capital. In addition to providing a measure of firm level productivity, as measured by the distance from the observation to the estimated production function, a production function can also provide industry–level economic insights such as returns to scale, elasticity of substitution between inputs, or most productive scale size. Thus, production functions are applied to many different fields including agriculture, banking, education, environment, health care, energy, manufacturing and so on (Fried et al. (2008)).

When modeling production functions, parametric regression models are still widely used although they require ex–ante specifications of a typically unknown functional form. Given that productivity is measured as the unobserved residual, overly restrictive assumptions about the production functions form are likely to result in a biased estimates of productivity levels. In contrast, nonparametric regression methods, such as the local linear estimator, avoid functional misspecification. However, the flexible nature of nonparametric methods make it difficult to interpret the estimation results in production economics (Beattie et al. (1985)). Fortunately, microeconomic theory such as Varian (1984) provides additional structures for modeling production which can be stated as 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 due to overfitting or computational feasibility. Thus, the first objective of this research is to propose a new estimator that imposes shape restrictions on local kernel weighting methods. By combining local averaging with shape-constrained estimation, we hope to improve finite sample performance by avoiding overfitting and control the computational complexity.

Furthermore, existing methods only allow to impose simple shape constraints such as concavity/convexity and monotonicity. These structure exclude important economic phenomena such as increasing returns to scale due to specialization, fixed costs, or learning (Frisch (1964)). This makes a shape constrained estimator biased and inconsistent regardless of a flexible nature of nonparametric estimator due to the misspecified constraints. This motivates us to impose more general flexible axiomatic structures consistent with the Regular Ultra Passum (RUP) law. There are several existing estimators which impose the RUP law as shape constraints. However, these estimation methods assume additional structure such as homotheticity on production function, and thus, are not flexible enough to capture a variety of realistic production structures. Thus, the objective of this research is to impose more general shape restrictions on a production function to allow the characterization of wider variety of economic phenomena.

Finally, when estimating production functions, endogeneity is a common critical problem since firms' managers determine input levels while knowing their firm specific productivity level which is modeled as a part of the residuals. Specifically, firms change variable labor input levels due to the shocks in productivity, which makes conventional regression models biased and inconsistent. For instance, more productive firms may hire more labors since firms need to prepare for a busy period in the near future. Several solutions to this endogeneity problem have been suggested including instrumental variable (IV) and control function approaches. However, the existing methods assume specific parametric functional forms on the production function, and thus, are likely to suffer from functional misspecification. Thus, we extend the proposing shape constrained estimator to the case when variable inputs are endogenous by using instrumental variables. We propose to impose shape constraints in the Landweber–Fridman regularization estimation process. This estimator is computationally feasible even with complicated shape constraints.

We estimate production functions for Chilean and Japanese manufacturing industries with our proposed models. The estimation results provide a description of supply–side of manufacturing industries as we report marginal product, marginal rate of substitution, elasticity of substitution and the most productive scale size. We find that these models result in a alternative economic insights than existing parametric models. Specifically, restrictive parametric models are likely to be suffered from the bias due to the misspecification. Furthermore, we report the aggregated productivity–level to measure the dynamics of industry growth over time.

The remainder of this dissertation is as follows. In Chapter 2, we propose a novel method called *Shape Constrained Kernel-weighted Least Squares* (SCKLS), which optimizes a local polynomial kernel criterion while estimating a multivariate regression function with simple shape constraints such as concavity and monotonicity. In 3, we generalize shape constraints and propose an iterative algorithm to estimate a production function which satisfies both the S-shape restriction and input isoquant convexity. In Chapter 4, we propose to extend our models using an IV approach to address the endogeneity by imposing shape constraints on a regularization process. In Chapter 5, we conclude and summarize the entire work, and suggest the research direction for future work.

2. SHAPE CONSTRAINED KERNEL-WEIGHTED LEAST SQUARES: ESTIMATING PRODUCTION FUNCTIONS FOR CHILEAN MANUFACTURING INDUSTRIES [1]

2.1 Introduction

In this chapter, we propose a new estimator that imposes shape restrictions on local kernel weighting methods. By combining local averaging with shape-constrained estimation, we hope to 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 chapter. 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

[1] Reprinted with permission from “Shape constrained kernel-weighted least squares: Estimating production functions for Chilean manufacturing industries. by Yagi, D., Chen, Y., Johnson, A. L., & Kuosmanen, T., 2018. *Journal of Business & Economic Statistics*, just-accepted.”, Copyright by Taylor & Francis.

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 chapter, *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 chapter, 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 [2] 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. Having 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

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

alternative shape restrictions, as discussed in Appendix A.1.

The remainder of this chapter is as follows. Section 2.2 describes the model framework and presents our estimator, SCKLS. Section 2.3 contains the statistical properties of the estimator, and Section 2.4 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 2.5. Section 2.6 applies the SCKLS estimator to estimate a production function for both the Chilean plastics and wood industries. Section 2.7 concludes and suggests future research directions. Appendix A.1 provides extensions to SCKLS and a comparison to CNLS and CWB. Appendix A.2 contains all the technical proofs and Appendix A.3 describes a test for affinity. Appendix A.4 states the details of the iterative algorithm for SCKLS, and Appendix A.5 presents a more extensive set of simulation results. Appendix A.6 describes the details of the partially linear model, and Appendix A.7 gives further details about the application to the Chilean manufacturing data.

2.2 Model Framework and Methodology

2.2.1 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 $g_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ 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.

2.2.2 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

$$\min_{a,\boldsymbol{b}} \sum_{j=1}^{n} (y_j - a - (\boldsymbol{X}_j - \boldsymbol{x})'\boldsymbol{b})^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}}{\boldsymbol{h}}\right), \tag{2.1}$$

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

$$\hat{g}_n(\boldsymbol{x}; \hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = \min_{i \in \{1,\ldots,m\}} \left\{ \hat{a}_i + (\boldsymbol{x} - \boldsymbol{x}_i)' \hat{\boldsymbol{b}}_i \right\} \tag{2.2}$$

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

$$\begin{aligned} \min_{\boldsymbol{a},\boldsymbol{b}} \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} (y_j - a_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)' \boldsymbol{b}_i)^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right) & \\ \text{subject to} \quad & a_i - a_l \geq \boldsymbol{b}_i'(\boldsymbol{x}_i - \boldsymbol{x}_l), & i, l = 1, \ldots, m \\ & \boldsymbol{b}_i \geq 0, & i = 1, \ldots, m. \end{aligned} \tag{2.3}$$

The first set of constraints in (2.3) 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 (2.2) 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 2.6 uses both a percentile grid and k-NN approach to define the kernel function. For details of these extensions, see Appendix A.1.

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 2.5 for the details.

Appendix A.1 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.

2.3 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 2.1.

(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,* $\sup_{\boldsymbol{x} \in \boldsymbol{S}} E\Big(\epsilon_j^4 \Big| \boldsymbol{X}_j = \boldsymbol{x}\Big) < \infty$.

(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 2.2.

(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 2.1(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 2.1. *Suppose that Assumption 2.1(i)-2.1(vi) and Assumption 2.2(i) or 2.2(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 2.2.

1. ***(The case of an increasing** m**)** Suppose that Assumption 2.1(i)-2.1(vi) and Assumption 2.2(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 2.1(i)-2.1(vi) and Assumption 2.2(ii) hold. Then, as $n \to \infty$, with probability one, the estimates from SCKLS satisfy*

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

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

Theorem 2.3. *Suppose that Assumption 2.1(i)-2.1(vi), Assumption 2.2(i) and Assumption 2.3 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.1 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.

2.4 Shape Misspecification: Theory and Testing

2.4.1 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 2.2(i)) and define $g^* : \boldsymbol{S} \to \mathbb{R}$ as

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

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 2.2.

Theorem 2.4.

Suppose that Assumption 2.1(i), 2.1(iii)-2.1(vi) and Assumption 2.2(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 2.4 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 A.5 for a numerical demonstration.

2.4.2 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

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

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

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

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 2.1(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[3]

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

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

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

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

[3] 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 $p_n = \frac{1}{B}\sum_{k=1}^{B} \mathbf{1}_{\{T_n \le T_{nk} + \Delta_n\}}$, 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.

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 A.2 for the proof.

Theorem 2.5. *Suppose that Assumptions 2.1(i),(iii)–(v) and 2.2(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 2.2(i) by Assumption 2.2(ii), the same conclusions hold for sufficiently large m.

See also Section 2.5 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 A.3 describes our procedure for testing affinity using SCKLS.

2.5 Simulation study

2.5.1 Numerical experiments on estimation

2.5.1.1 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 A.5 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 A.5 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[4]. We replicate each scenario 10 times and report the average and standard deviation.

2.5.1.2 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 relatively insensitive to the bandwidth choice (see for example Section 2.5.1.3.1). For further computational improvements, we apply the iterative algorithm described in Appendix A.4.

2.5.1.3 Results

2.5.1.3.1 Fixed number of evaluation points

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

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 2.1 and 2.2 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 A.5. 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 (2.3)). 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

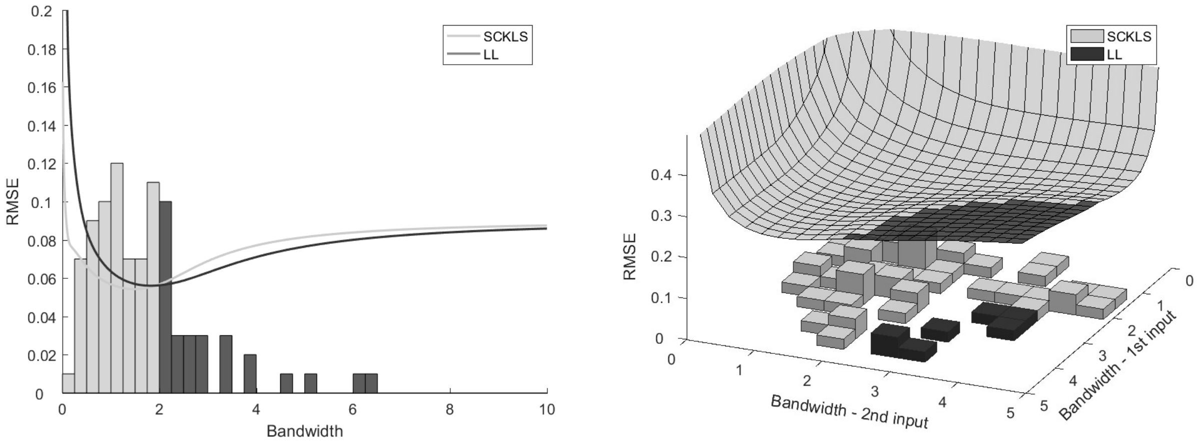
Table 2.1. RMSE on observation points for Experiment 1.

		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 2.2. RMSE on evaluation points for Experiment 1.

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

each bandwidth as well as for the bandwidth via LOOCV. Figure 2.1 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 - 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 A.5 for more details.



(a) One-input

(b) Two-input

Figure 2.1. The histogram shows the distribution of bandwidths selected by LOOCV. The curves show the relative performance of each estimator.

2.5.1.3.2 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 2.3 and 2.4 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 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 2.3. RMSE on observation points for Experiment 2.

		Average of RMSE on observation points		
Number of evaluation 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 2.4. RMSE on evaluation points for Experiment 2.

		Average of RMSE on evaluation points		
Number of 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

2.5.2 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:

$$g_0(x) = x^p \tag{2.4}$$

and

$$g_0(x) = \frac{1}{1 + \exp(-5\log(2x))}. \tag{2.5}$$

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 (2.4) 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 (2.5) 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 2.4.2 with $B = 200$.

Table 2.5 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 2.5. Rejection rate (%) of the test for monotonicity and concavity

Sample size	DGP Scenario	Power of the Test (α)			
		0.05	0.01	0.05	0.01
(n)		$\sigma = 0.1$		$\sigma = 0.2$	
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

2.6 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[5]. See Appendix A.7 for the details of estimation and comparison across different estimators.

2.6.1 The census of Chilean manufacturing plants

We use the Chilean Annual Industrial Survey provided by Chile's National Institute of Statistics[6]. 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[7]. Capital and Value Added are measured in millions of Chilean 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:

$$y_j = \boldsymbol{Z}_j' \boldsymbol{\gamma} + g_0(\boldsymbol{X}_j) + \epsilon_j, \tag{2.6}$$

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 establish-

[5] 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.

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

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

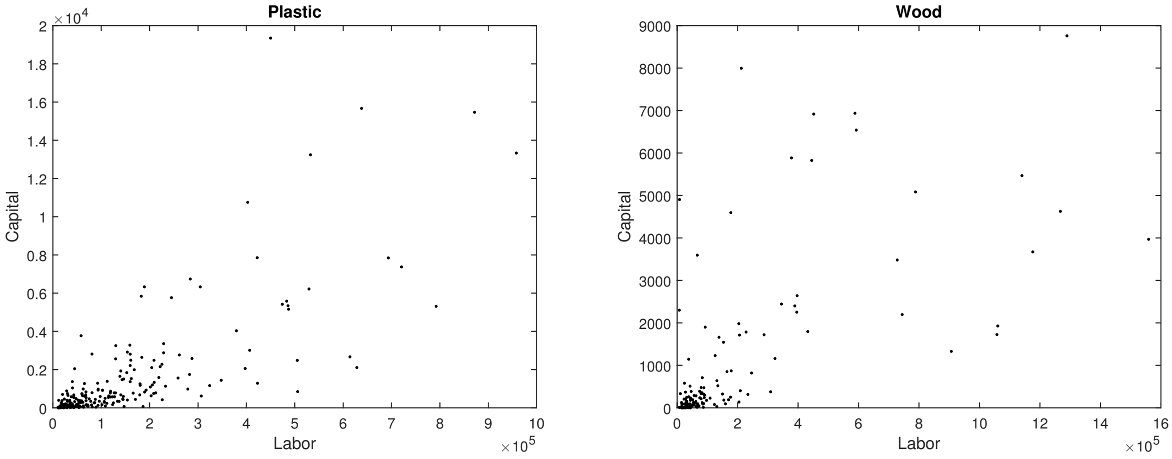
Table 2.6. Statistics of Chilean manufacturing data.

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

ments that are exporting and the share of output being exported. For more details see Appendix A.6.

Table 2.6 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 A.5), we use these options. We choose the smoothing parameter k via leave-one-out cross validation. Appendix A.1 explains the details of our implementation of K-NN for the SCKLS estimator.

Figure 2.2 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 A.7 for details.



(a) Plastic (2520)

(b) Wood (2010)

Figure 2.2. Labor and Capital of each industry.

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 (2015)[8]. 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 2.4.2 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.

[8] We apply a Cobb–Douglas OLS to the second stage data $\{\boldsymbol{X}_j, y_j - \boldsymbol{Z}_j\gamma\}_{j=1}^{n}$ which removes the effect of contextual variables from observed output. See Appendix A.6 for details.

2.6.2 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 2.7 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 2.7. SCKLS fitting statistics for cross sectional data.

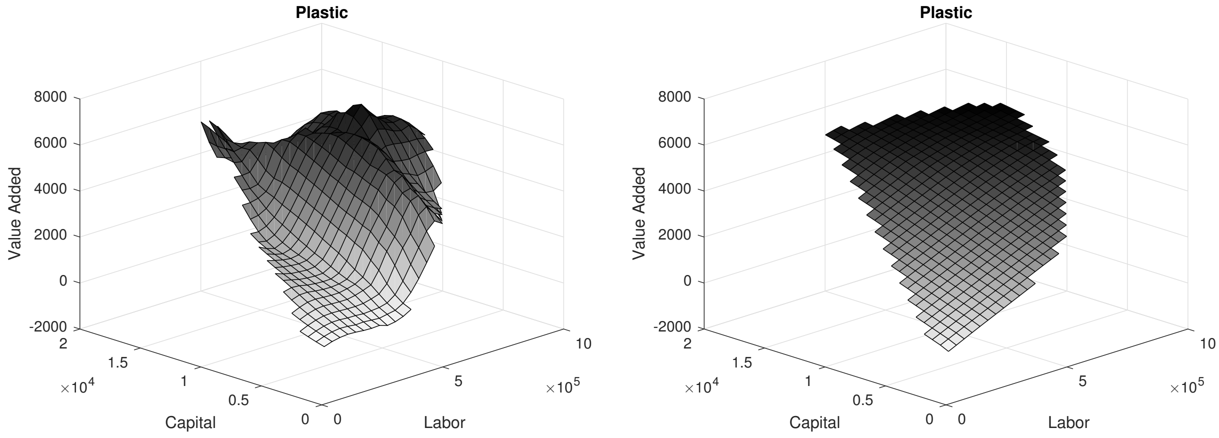
Industry	Number of observations	R^2
Plastic	245	71.1%
Wood	132	43.8%

Table 2.8 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 2.3 and Figure 2.4 show the estimated production function within the convex hull of observations for plastic and wood industries, respectively. Visually, the production 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 2.8. Characteristics of the production function.

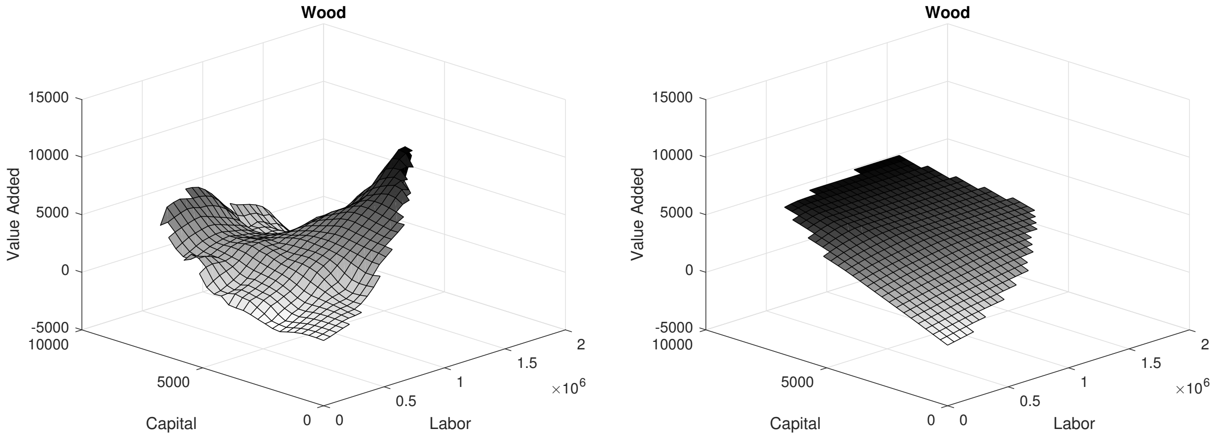
	Plastic (2520)		
	Marginal Productivity		Marginal Rate of Substitution
	Labor ($= b_l$) (million peso/man hours)	Capital ($= b_k$) (peso/peso)	($= b_k/b_l$)
10th percentile	0.00396	0.111	23.3
25th percentile	0.00523	0.139	23.9
50th percentile	0.00579	0.139	24.0
75th percentile	0.00579	0.139	35.3
90th 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}



(a) Local Linear

(b) SCKLS

Figure 2.3. Production function estimated by LL and SCKLS for the plastic industry (2520)



(a) Local Linear

(b) SCKLS

Figure 2.4. Production function estimated by LL and SCKLS shape constraints for the wood industry (2010)

Table 2.9 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 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 2.9. 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 2.10 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 2.10. 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

2.7 Conclusion

This chapter 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.

3. AN AXIOMATIC NONPARAMETRIC PRODUCTION FUNCTION: MODELING PRODUCTION IN JAPAN'S CARDBOARD INDUSTRY

3.1 Introduction

The goal of this chapter is to develop a new approach that is less dependent on functional form assumptions to estimate a production function. Our basic idea is to use nonparametric local averaging methods, but augment these methods with shape constraints that reflect economic axioms. Nonparametric local averaging methods without shape constraints would avoid the potential for functional form misspecification and flexibly capture the nuances of the data, but would be difficult to interpret economically and would not satisfy the basic properties, i.e. globally convex input isoquants or a well-defined marginal product estimates. Thus, we can use a minimal set of economic axioms which are unlikely to be violated while providing additional structure. The axioms we impose are the Regular Ultra Passum (RUP) law as the scaling property[1] and that input isoquants are both convex and non-homothetic. This new modeling approach estimates the most productive scale size conditional on input mix.

The RUP law states that along any expansion path, the production function should first have increasing returns-to-scale followed by decreasing returns-to-scale, Frisch (1964). Intuitively, when a firms is small it tends to face increasing returns-to-scale because it can increase productivity easily through specialization and learning. In contrast, as the scale size becomes larger, a firm tends to have decreasing returns-to-scale due to scarcity of ideal production inputs and challenges related to increasing span of control. Firms in competitive markets should operate close to the most

[1] As explained below, we will actually use an S-shape restriction which requires a single inflection point, but otherwise generalizes the RUP law.

productive scale size in the long-run to minimize the cost per unit and assure positive profits. The RUP law will assure we have a well-defined marginal products and most productive scale sizes, Frisch (1964).

Convex input isoquants, which are a standard assumption in production theory, are motivated by the argument that there are optimal proportions in which inputs should be used for production and that deviations from the optimal proportion by decreasing the level of one input, such as capital, will require more than a proportional increase in another input, such as labor. Relaxing the homotheticity of input isoquants allows the optimal proportions to depend on the output level. For example, the optimal proportion of inputs for low output levels could be more labor intensive than at higher output levels. Further, non-homothetic isoquants allows for the most productive scale size measured along different rays from the original to exist at different output levels. Non-homothetic isoquants allows us to more easily capture the empirical fact that productivity levels are a function of capital intensity.

The axiomatic approach is critical for interpreting the estimates of a production function to gain managerial insights. The production function is often used to estimate firm expansion behavior including how many resources need to be added to expand output or how automation (i.e. changing the capital-to-labor ratio) can be used to achieve larger scales of production. Without data and production function estimates, managers are left to make these decisions based on a firms historical behavior or rules-of-thumb or other approximations. The analysis of firm as a whole allows for the accounting of synergies between inputs in the production process.

We implement our approach using data from Japan's corrugated cardboard industry. As classified in the Japanese Census of Manufactures, the cardboard industry[2] includes both cardboard

[2] In the *Japan Standard Industrial Classification* (JSIC) the corrugated cardboard industry is industry (1453).

manufacturers and cardboard box manufacturers. The latter sector is not particularly capital intensive nor does it require technical know-how to enter, thus firms focus on customer service and supplying slightly customized products to maintain market share. Overall, the industry has a few large firms and many smaller firms, which is a typical structure for a mature manufacturing industry. The largest firms in the industry are vertically integrated and include cardboard production, box making, and paper making.[3]

In the cardboard industry, like most industries, firms enter the market as small firms and must expand over time taking advantage of capital and labor specialization or other characteristics of the technology to be more productive, Haltiwanger et al. (2013); Foster et al. (2016). Recently, medium and large sized firms in the industry have been acquiring smaller firms and reducing the combined input levels without significant reductions in the combined output levels, leading to higher productivity levels. In particular, since the medium and small size firms are operating below the most productive scale size, they have the potential for significant increase in productivity by increasing their scale of production, thus mergers are attractive to medium sized firms.

Several nonparametric shape constrained estimators have been proposed that combine the advantage of avoiding functional misspecification with improving the interpretability of estimation results relative to unconstrained nonparametric methods, see for example Kuosmanen et al. (2015) or Yagi et al. (2018). However, existing methods only allow the imposition of simple shape constraints such as concavity and monotonicity. These structures exclude economic phenomena such as increasing returns to scale due to specialization, fixed costs, or learning. Thus, more general functional structures, like the model proposed in this chapter, are desirable.

[3] In the Census of Manufacturing, establishments are classified by industry based on the primary product produced in the establishment. Paper making establishments are typically specialized and do not appear in our data set. However, vertically integrated firms that own paper producing establishments typically have larger cardboard and box making establishments.

There have been two previous attempts to develop estimators that impose the RUP law as shape constraints. The first, Olesen and Ruggiero (2014) develops an algorithm to estimate a Data Envelopment Analysis (DEA)-type estimator satisfying the RUP law and impose homotheticity on the input isoquants. Noise is not modeled in DEA estimators and all deviations from the estimated function are one-sided and negative. Hwangbo et al. (2015) introduce noise and estimate a scaling function using nonparametric shape constrained methods. However, they also assume homothetic input isoquants and do not provide statistical properties for their estimators. In conclusion, these estimation methods place structure on production function, but the homothetic assumption is not flexible enough to capture a variety of realistic and potential production structures.

We will use our production model to provide a description of the supply-side of the Japanese cardboard industry as we report marginal product, marginal rate of substitution, most productive scale size and productivity evolution. We find most productive scale size is dependent on the capital-to-labor input factor ratio and the largest firms operate close to the largest most productive scale size associated with a high capital-to-labor ratio. However, we also find that firms' productivity is negatively correlated with capital-to-labor ratio, indicating firms improve productivity by reducing capital-to-labor ratio.

We also look at productivity variation. The standard approach is to calculate productivity uses factor elasticities as weights to aggregate the various inputs. Syverson (2004) summarizes this approach and uses input cost shares of the individual firms to approximate the factor elasticities. He reports that within four-digit Standard Industry Codes (SIC) industries in the U.S. manufacturing sector, the average difference in total factor productivity (TFP) between an industry's 90th and 10th percentile firms implies the firm at the 90th percentile of the productivity distribution makes almost twice as much output with the same measured inputs as the 10th percentile firm. Hsieh and

Klenow (2009) finds even larger productivity differences in China and India, with average 90-10 TFP ratios over 5:1.

We study the productivity variation in the Japanese cardboard industry and find the 90-10 TFP ratio is 4:1. Using our approach to estimating production functions, we find the production function can explain approximately 25% of the productivity variation in the industry. This means that because of specialization, fixed costs, and learning we expect smaller firms to have lower productivity and by allowing non-homothetic input isoquants we characterize differences in substitution rates between inputs at different output levels. These scale and mix effects account for significant portion of the observed productivity variation leaving a much smaller component of unexplained productivity variation.

The remainder of this chapter is as follows. Section 3.2 introduces the proposed production function model and its assumptions. Section 3.3 and Appendix B.2 explain the two-step estimation procedure and the algorithm for our estimator, respectively. Section 3.4 discusses the Monte Carlo simulation results under several different experimental settings. Section 3.5 applies our estimator to estimate a production function for the Japanese cardboard industry. Section 3.6 concludes and suggests future research directions.

3.2 Model

Consider the following production function model

$$y = g_0(\boldsymbol{x}), \tag{3.1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)'$ is d-dimensional input vector, y is an output scalar, and $g_0 : \mathbb{R}^d_+ \to \mathbb{R}_+$ is a production function. We can rewrite this function as

$$\phi(y, \boldsymbol{x}) = y - g_0(\boldsymbol{x}). \tag{3.2}$$

Here we define mathematically an input isoquant,

Definition 3.1. *An input isoquant $\bar{V}(y) = \{\boldsymbol{x} : g_0(\boldsymbol{x}) = y\}$ be the sets of input vectors capable of producing each output vector y.*

We make the following assumptions on g_0 and ϕ:

Assumption 3.1.

(i) $g_0(\cdot)$ *is a strictly monotonically increasing function defined on a compact set.*

(ii) $\phi(\cdot, \cdot)$ *is a twice-differentiable function.*

(iii) $\frac{\partial \phi}{\partial x_k} = \frac{\partial g_0}{\partial x_k} \in (0, \infty)$ *for every* $k = 1, \ldots, d$ *over the domain of* g_0.

Under Assumption 3.1, by the implicit function theorem, there exists an implicit function $\mathcal{H}_k$ such that

$$x_k = \mathcal{H}_k(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d; y) = \mathcal{H}_k(\boldsymbol{x}_{-k}; y) \quad \forall k = 1, \ldots, d, \tag{3.3}$$

where $\boldsymbol{x}_{-k} = \{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d\}$ is an input vector without the k-th input.

We are interested in estimating a production function g_0 having both convex input isoquants for all output levels and that satisfies an augmented version of the RUP law. We assume input convexity which implies the following conditions on $\mathcal{H}_k$:

Definition 3.2. *An input isoquant is input-convex if for any given value for an arbitrary input x_k, then for every pair of arbitrary input vectors $\boldsymbol{x_a}, \boldsymbol{x_b} \in \mathbb{R}^{d-1}$, $y \in \mathbb{R}_+$ and $\lambda \in [0, 1]$,*

(i) $\lambda \mathcal{H}_k(\boldsymbol{x_a}; y) + (1 - \lambda)\mathcal{H}_k(\boldsymbol{x_b}; y) \geq \mathcal{H}_k(\lambda \boldsymbol{x_a} + (1 - \lambda)\boldsymbol{x_b}; y)$ *(Convex input isoquant),*

(ii) If $\boldsymbol{x_a} \leq \boldsymbol{x_b}$, then $\mathcal{H}_k(\boldsymbol{x_a}; y) \geq \mathcal{H}_k(\boldsymbol{x_b}; y)$ (Monotone decreasing input isoquant).

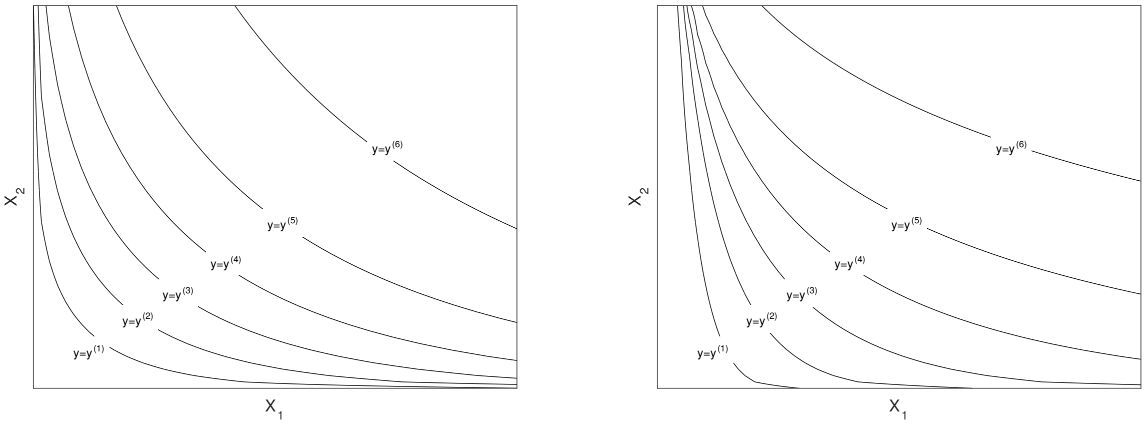
Intuitively, input convexity implies the existence of an optimal ratio of inputs. Deviations from the optimal input ratios by decreasing the use of a particular input will result in more than a proportional increase in other inputs. Further, larger deviations from the optimal ratio will require larger increases in input consumption to maintain the same output level.

Another common assumption for production functions is homotheticity.

Definition 3.3. *A production function g_0 is homothetic if for every $\boldsymbol{x}$ and $\alpha > 0$, the implicit function $\mathcal{H}_k$ is homogeneous of degree one*

$$\alpha x_k = \mathcal{H}_k(\alpha x_1, \ldots, \alpha x_{k-1}, \alpha x_{k+1}, \ldots, \alpha x_d; g_0(\alpha \boldsymbol{x})) \quad \forall k = 1, \ldots, d.$$

Input homotheticity is a strong assumption because it restricts input elasticity to be constant for a given input mix at all scales of production. However, by relaxing input homotheticity and assuming only input-convexity, each isoquant can have different curvatures at a given y-level. We refer to isoquants of this type as non-homothetic, convex input isoquants. Figure 3.1 shows homothetic and non-homothetic isoquants for a two dimensional input vector.



(a) Homothetic

(b) Non-homothetic

Figure 3.1. Input isoquants satisfying input convexity.

Next, we define the elasticity of scale[4], $\epsilon(\boldsymbol{x})$, relative to a production function $g_0(\boldsymbol{x})$:

$$\epsilon(\boldsymbol{x}) = \sum_{k=1}^{d} \frac{\partial g_0(\boldsymbol{x})}{\partial x_k} \frac{x_k}{g_0(\boldsymbol{x})}. \tag{3.4}$$

The Regular Ultra Passum (RUP) law is defined as follows:

Definition 3.4. *(Førsund and Hjalmarsson (2004)) A production function $g_0(\boldsymbol{x})$ obeys the Regular Ultra Passum law if $\frac{\partial \epsilon(\boldsymbol{x})}{\partial x_k} < 0$ for $\forall k = 1, \ldots, d$, and for some input $\boldsymbol{x_a}$ we have $\epsilon(\boldsymbol{x_a}) > 1$, and for some input $\boldsymbol{x_b}$ we have $\epsilon(\boldsymbol{x_b}) < 1$, where $\boldsymbol{x_b} > \boldsymbol{x_a}$.*[5,6]

Intuitively, for any ray from the origin, a production function g_0 has increasing returns to scale

[4] This variable was referred to as the passum coefficient in the seminal work of Frisch (1964), but is now commonly referred to as the elasticity of scale.

[5] $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are vectors such that the inequality implies that every component of $\boldsymbol{x_b}$ is greater than or equal to every component of $\boldsymbol{x_a}$.

[6] Note this definition of the RUP law generalizes Frisch (1964) the original definition. This definition does not require the passum coefficient to drop below 0, thus implying congestion or that the production function is not monotonically increasing. This characterization allows for a monotonically increasing production function. Also note that although a concave production function nests within this definition,the definition does not require that the function is "nicely concave" as defined in Ginsberg (1974).

followed by decreasing returns to scale. However, note that in both Førsund and Hjalmarsson (2004) and Frisch's original definition, neither rules out the possibility of multiple inflection points, see Appendix B.4 for a more detailed explanation. Furthermore, because the RUP law is defined in terms of the elasticity of scale, the law does not allow the function, g_0, to grow at an exponential rate. Thus, we introduce the following definition of an S-shape function.

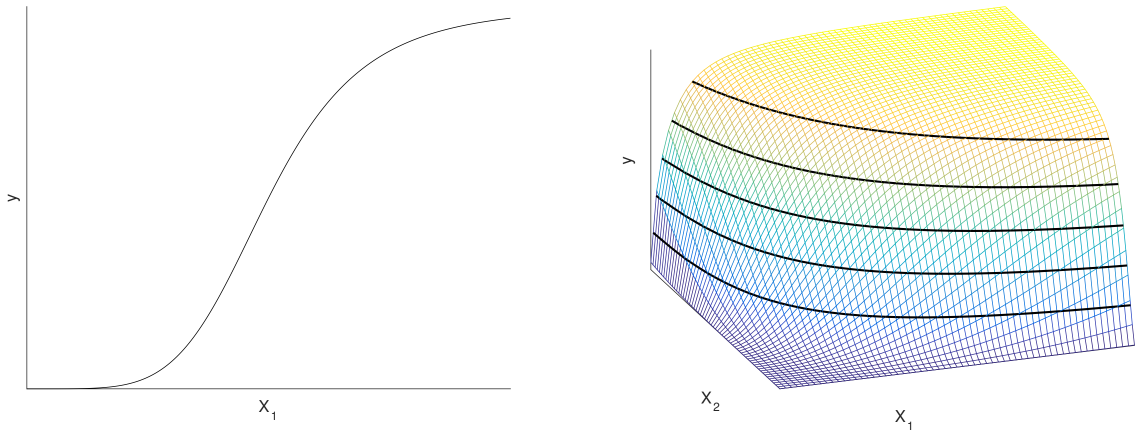
Definition 3.5. *A production function $g_0 : \mathbb{R}^d \to \mathbb{R}$ is S-shaped if for any $\boldsymbol{v} \in \mathbb{R}^d_+$ defining a ray from the origin in input space $\alpha\boldsymbol{v}$ with $\alpha > 0$, $\nabla^2_v g_0(\alpha\boldsymbol{v}) > 0$ for $\alpha\boldsymbol{v} < \boldsymbol{x}^*$, and $\nabla^2_v g_0(\alpha\boldsymbol{v}) < 0$ for $\alpha\boldsymbol{v} > \boldsymbol{x}^*$ along a ray from the origin, where $\nabla^2_v g_0$ is the directional second derivative of g_0 along v. This implies that for any ray from the origin of direction v, there exists a single inflection point $\boldsymbol{x}^*$ that $\nabla^2_v g_0(\boldsymbol{x^*}) = 0$.*

Given the definition of an S-shape function, the following lemma defines formally the relationship between the RUP law and an S-shape function. Proof is provided in Appendix B.1.

Lemma 3.1. *If a production function $g_0 : \mathbb{R}^d \to \mathbb{R}$ is second-differentiable, monotonically increasing and satisfies the regular ultra passum law and there exists a single inflection point x^* where $\nabla^2_v g_0(\boldsymbol{x^*}) = 0$ for any ray from the origin defined by a direction $\boldsymbol{v} \in \mathbb{R}^d_+$, then g_0 is S-shaped.*

Figure 3.2 (a) and (b) show two examples of the production function with one-input and two-input, respectively. Both functions satisfy the RUP law and the S-shaped definition. Furthermore, the two-input example (i.e. $d = 2$) has convex input sets.

In the following, we prove that a homothetic production function which satisfies the S-shape definition for a single ray from the origin will also satisfy the S-shape definition for any expansion path. To achieve this, we require the following alternative characterization for a homothetic production function. Proofs of all the theorems are deferred to Appendix B.1.



(a) one-input

(b) two-input

Figure 3.2. Production functions satisfying both the RUP law and S-shape definition

Definition 3.6. *(Alternative definition of homothetic production function) A production function $g_0(\boldsymbol{x}) = F(H(\boldsymbol{x}))$ is homothetic if*

(i) Scale function $F(\cdot)$ is a strictly monotone increasing function, and

(ii) Core function $H(\cdot)$ is a homogeneous of degree 1 function which implies $h(t\boldsymbol{x}) = tH(\boldsymbol{x})$ for all $t > 0$.

Define $X_{max,k} = \max \boldsymbol{X}_k$ for $\forall k = 1, \dots, d$ and $\boldsymbol{X}_M = (X_{max,1}, ..., X_{max,k}, ..., X_{max,d})$. And also define $\boldsymbol{X}_0 = \boldsymbol{0}$. The value of the core function, g when evaluating the input vector, $\boldsymbol{X}$, is referred to as aggregate input, specifically $x_A = H(\boldsymbol{X})$.

Definition 3.7. *A rising curve (commonly referred to as an expansion path), $\boldsymbol{EP}$, is a series of $M+1$ input vectors, $\{\boldsymbol{X}_0, \dots, \boldsymbol{X}_M\}$ such that $x_{A,m} < x_{A,m+1}$ for every $m = 0, \dots, M-1$, where $x_{A,m} = H(\boldsymbol{X}_m)$. In addition, we denote $\Big\{\big(H(\boldsymbol{X}_0), g_0(\boldsymbol{X}_0)\big), \dots, \big(H(\boldsymbol{X}_M), g_0(\boldsymbol{X}_M)\big)\Big\}$ as the* core *of $\boldsymbol{EP}$.*

Lemma 3.2. *Assume a production function is homothetic in inputs and the S-shape definition holds for one particular ray from the origin. Consider any pair of rays from the origin and define two 2-D sectionals of the production function. For both rays from the origin, the S-shape definition is satisfied and the inflection points lie on the same input isoquant with aggregate input level, x_A^*.*

Theorem 3.1. *Assume a production function is homothetic in inputs. If the S-shape definition holds for a single ray from the origin, then the S-shape definition will hold for the core of any expansion path.*

3.3 Estimation Algorithm

3.3.1 Framework

Given observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ satisfying $y_j = g_0(\boldsymbol{X}_j) + \epsilon_j$, where ϵ_j are i.i.d. noise with zero-mean and finite variance. Our goals include the following:

1. For a given level y, estimate the isoquant function satisfying both the convex input and the monotone decreasing input assumptions (see Definition 3.2).

2. For a given direction $\boldsymbol{v} \in \mathbb{R}_+^d$, estimate the production curve, i.e. $g_0(\alpha \boldsymbol{v})$ for $\alpha > 0$, satisfying monotonicity and S-shaped assumptions (see Definition 3.5).

3. Given the unit cost of each input as well as the total budget, tackle problems based on our estimators of the quantities mentioned above, such as optimal resource allocation.

3.3.2 Overview

We propose an estimation algorithm for a production function satisfying both the S-shape definition and input convexity without any further structural assumptions. The algorithm combines two different shape constrained nonparametric estimation methods. Succinctly, the algorithm is constructed by two estimations: (1) Input isoquants for a set of y–levels, and (2) S-shape functions on a set of rays from the origin. Algorithm 1 presents our basic algorithm which is composed of these two estimators.[7] We reference a pilot estimate which can be any estimator that will provide an initial rough estimate of the function. The right-hand column of Algorithm 1 reports the section numbers where the details of each step are described.

We approximate a production function g_0 with isoquant estimates for a set of output levels, and S-shape functional estimates for a set of rays from the origin as shown in Figure 3.3(a). We also develop the interpolation procedure to obtain the functional estimates $\hat{g}_0(\boldsymbol{x})$ at any given input $\boldsymbol{x}$. Figure 3.3(b) shows the interpolated surface of the estimated production function.

Algorithm 1 Basic estimation algorithm

```
Data: observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$
procedure                                                                  (Section)
Initialization:                                                            (3.3.3)
    $I \leftarrow$ Initialize number of isoquants
    $R \leftarrow$ Initialize number of rays
    $\{y^{(i)}\}_{i=1}^I \leftarrow$ Initialize isoquant $y$-levels with $y^{(1)} < \cdots < y^{(I)}$.
    $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^R \leftarrow$ Initialize rays from origin
Estimation:                                                                (3.3.4)
    For $j = 1, \ldots, n$, let $\tilde{y}_j = \tilde{g}_0(\boldsymbol{X}_j)$, where $\tilde{g}_0$ is the pilot estimator of $g_0$
    Project $\{\boldsymbol{X}_j, \tilde{y}_j\}_{j=1}^n$ to the isoquant level $y^{(i)}$
    Estimate convex isoquants by the CNLS-based estimation
    Project observations onto the ray $\boldsymbol{\theta}^{(r)}$
    Estimate S-shape functions using the SCKLS-based estimator
return : Estimated function with minimum Mean Squared Errors
```

Since we estimate the S-shape function on rays from the origin, it is convenient to use a spher-

[7] The algorithm refers to CNLS-based and SCKLS-based estimators for a description of these methods see Appendix B.2.2 and Appendix B.2.3.2 respectively.

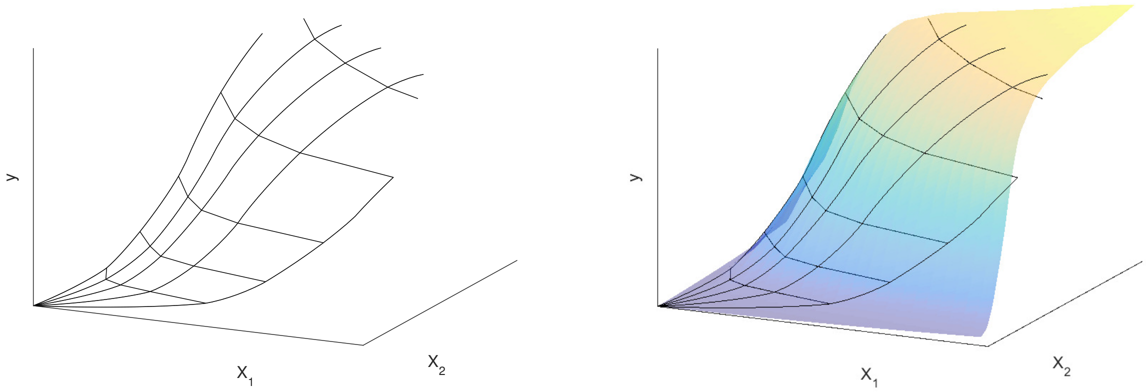
ical coordinates system which is defined by the angle and distance (radius) of observed points to the origin. Therefore, our observed input vector $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jd})'$ in spherical coordinates system $(r_j, \boldsymbol{\phi}_j) = (r_j, \phi_{j,1}, \ldots, \phi_{j,d-1})$ is defined as:

$$\begin{aligned}
r_j &= \sqrt{X_{j1}^2 + \ldots + X_{jd}^2} \\
\phi_{j,1} &= \arccos \frac{X_{j1}}{\sqrt{X_{j1}^2 + \ldots + X_{jd}^2}} \\
\phi_{j,2} &= \arccos \frac{X_{j2}}{\sqrt{X_{j2}^2 + \ldots + X_{jd}^2}} \\
&\vdots \\
\phi_{j,d-2} &= \arccos \frac{X_{j,d-2}}{\sqrt{X_{j,d-2}^2 + X_{j,d-1}^2 + X_{jd}^2}} \\
\phi_{j,d-1} &= \arccos \frac{X_{j,d-1}}{\sqrt{X_{j,d-1}^2 + X_{jd}^2}},
\end{aligned} \tag{3.5}$$

where r_j is the radial distance from the origin, and $\{\phi_{j,1} \ldots \phi_{j,d-1}\}$ defines the angle of the observation.

3.3.3 Initialization

We initialize the parameters used in the estimation. The number of isoquants I and the number of rays from the origin R affect the flexibility of the estimated function (computation time increases with the number of isoquants and rays). We initialize isoquant y-levels, $\{y^{(i)}\}_{i=1}^{I}$, and rays from the origin, $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^{R}$, based on the distribution of the observations. We propose three options: (1) Evenly spaced grid, (2) Equally spaced percentile grid, and (3) Centroid of K-means cluster of observations. To set notation, given the number of isoquants, I, and rays, R, we set the grid as $y^{(i)}$ and $\boldsymbol{\theta}^{(r)}$, the locations of the isoquants and rays respectively. To overcome skewness in the



(a) Functional estimates

(b) Interpolated functional estimates

Figure 3.3. Illustration of functional estimates.

empirical data in which there are many smaller firms and only a few large firms, we recommend an equally spaced percentile grid or K-means cluster .

3.3.4 Two–step estimation

During the estimation step, we approximate the production function by estimating the isoquants at a set of y-levels and estimating the S-shape functions on a set of rays from the origin. We calculate the estimates over different tuning parameters, compute the mean squared errors (MSE) against observations, and return the final estimates corresponding to the tuning parameters with the minimum MSE.

3.3.4.1 Isoquant estimation

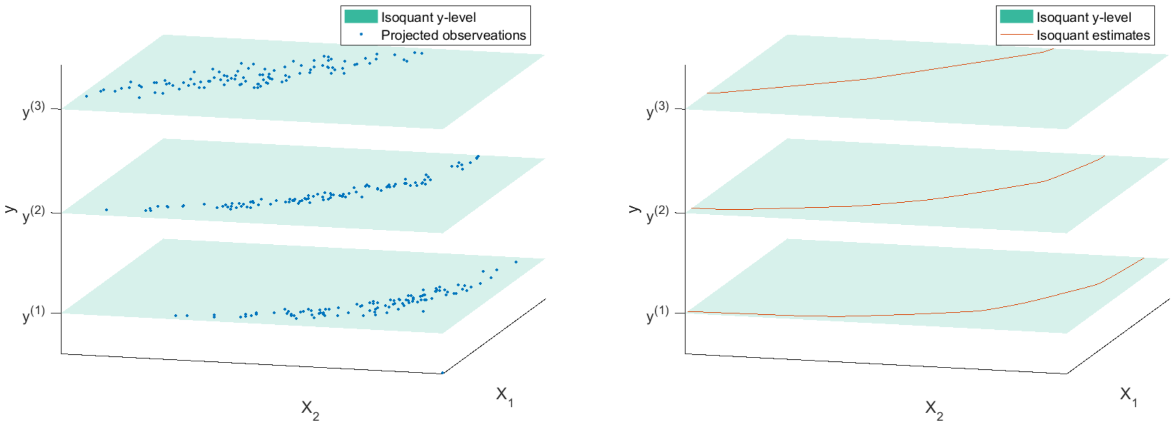
Before estimating the isoquants, we need to assign each observation $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ to an isoquant y-level, $y^{(i)}$ based on $\tilde{y}_j$ from a pilot estimator. The purpose of the pilot estimator is to improve the classification of observations to isoquant levels. Most well-known nonparametric estimators, such as local linear estimator could be used. We suggest simply assigning each observation to the closest isoquant y-level, which means

$$i_j = \underset{i \in \{1,\ldots,I\}}{\operatorname{argmin}} \left(\tilde{y}_j - y^{(i)}\right)^2 \quad \forall j = 1, \ldots, n, \tag{3.6}$$

where i_j indicates the isoquant index to which we assign observation j. Then, we define the projected observations for the i^{th} isoquant as $\{\boldsymbol{X}_j, y^{(i_j)}\}_{\{j:i_j=i\}}$, where $y^{(i_j)}$ is the output level of the i^{th} isoquant. Figure 3.4(a) shows the projection of each observation to the corresponding isoquant y-level. We estimate a set of isoquants using the CNLS-based method which is a nonparametric estimation method imposing convexity for each $y^{(i)}$-level. Intuitively, we estimate the convex isoquant estimates nonparametrically without imposing any ex ante functional specification for each $y^{(i)}$-level. Figure 3.4(b) shows the isoquant estimates obtained with projected observations $\{\boldsymbol{X}_j, y^{(i_j)}\}$. The mathematical formulation is described in Appendix B.2.2. Finally, for the estimation of an isoquant at any y-level (with $y \in [\min_i y^{(i)}, \max_i y^{(i)}]$), we first select the two closet isoquants associated with a larger and smaller output level in $\{y^{(i)}\}_{i=1}^{I}$, namely, $y^{(i^+)}$ and $y^{(i^-)}$, and then return the convex combination of these two isoquants with the weights $\frac{y-y^{(i^-)}}{(y^{(i^+)}-y^{(i^-)})}$ and $\frac{y^{(i^+)}-y}{(y^{(i^+)}-y^{(i^-)})}$ respectively.

3.3.4.2 S-shape estimation

To estimate the S-shape functions on rays from the origin, we begin by project all observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^{n}$ to each ray from the origin $\boldsymbol{\theta}^{(r)}$. We use the estimated isoquants from the previous step to project the observations. In short, we find the level of an isoquant that $\boldsymbol{X}_j$ belongs to it. Below we also provide an alternative way of thinking about this step. Considering the observations input level, $\boldsymbol{X}_j$, we select the two closest isoquants associated with a larger and smaller aggregate inputs. Here the definition of larger and smaller vectors are in terms of a proportional expansion



(a) Projected observations to each $y^{(i)}$

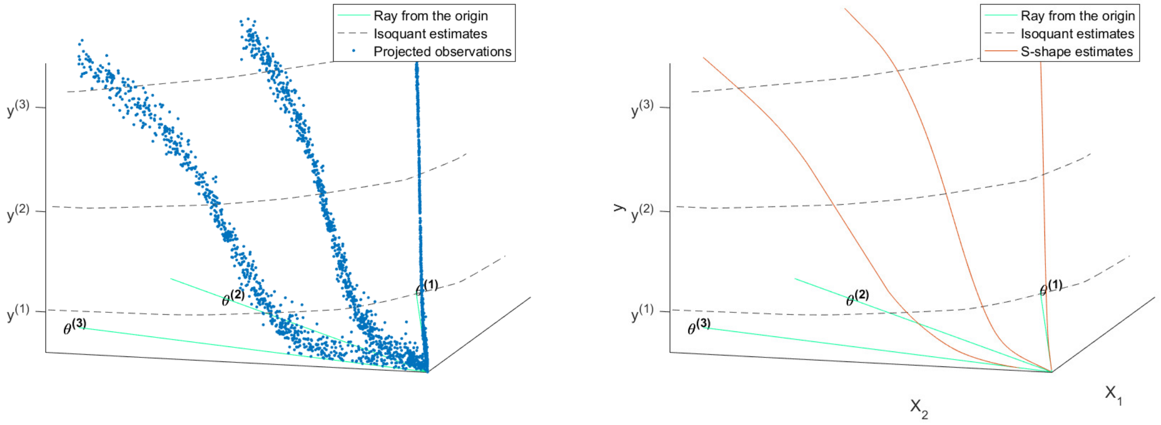
(b) Isoquant estimates on $y^{(i)}$

Figure 3.4. Isoquant estimation

or contraction of the input vector, $\lambda \boldsymbol{X}_j$ where $0 \leq \lambda < \infty$ with $\lambda \geq 1$ indicating expansion and $\lambda \leq 1$ indicating contraction. We will refer the two closest isoquants as “sandwiching" the input vector of interest. Then, we assign weights to these two isoquants based on the distance to the observed input $\boldsymbol{X}_j$ along a ray from the origin through the observed points. Finally, we project the observation with the weighted average of the two isoquant estimates. Figure 3.5(a) shows the projection of our observations. The details are described in Appendix B.2.3.1.

Next, we use the SCKLS-based method to estimate the S-shape function on each ray from the origin. Note that this estimation assigns two different kernel weights to each observation. The first weight is a function of the angle(s) formed by a ray from the origin through the observation and a ray from the origin through the current evaluation point. This will be a vector of bandwidths if there are more than two regressors. The second weight is a function of the distance measured along the ray between the projected observation and the evaluation point.

SCKLS-based estimation requires the selection of a smoothing parameter which we refer to as the bandwidth. Intuitively, a smaller bandwidth will lead to over–fitting the data, and a larger bandwidth will lead to over–smoothing. Thus, it is crucial to select the optimal bandwidth by balancing the bias–variance tradeoff of the estimator. In our algorithm, the bandwidth of the kernel weights for angles, $\boldsymbol{\omega}$, is optimized via a grid search, and the bandwidth of the kernel weights for distance along the ray, $h^{(r)}$, is optimized by leave-one-out cross-validation, given kernel weights for angles. Figure 3.5(b) shows the S-shape estimates obtained with projected observations. The mathematical details are described in Appendix B.2.3.2.



(a) Projected observations to each $\boldsymbol{\theta}^{(r)}$

(b) S-shape estimates on $\boldsymbol{\theta}^{(r)}$

Figure 3.5. S-shape estimation

3.3.4.3 Computing functional estimates at a given input vector

The last step of Algorithm 1 obtains the functional estimates $\hat{g}(\boldsymbol{x})$ at any given value of input vector $\boldsymbol{x}$, and computes the MSE against observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$.

First we compute the weighted average of the two closest isoquants which sandwich the observed input $\boldsymbol{X}_j$. The details are given in Appendix B.2.3.1. Second, we assign weights to each

S-shape estimate based on the angle between a given input vector $\boldsymbol{x}$ and each ray from the origin $\boldsymbol{\theta}^{(r)}$ on which we have estimated the S-shape functions followed by computing the weighted average of the S-shape estimates and obtaining the final functional estimates on a given input $\boldsymbol{x}$, $\hat{g}(\boldsymbol{x})$. Figure 3.3(b) shows the interpolated functional estimates. The details are given in Appendix B.2.3.5.

Note that there may be a gap between the convex isoquant estimates and the S-shape estimates on rays from the origin. Specifically, if the S-shape estimates do not all lie on the input isoquant for each evaluated output level $y^{(i)}$, then the S-shape estimates will not match the isoquant estimates at some isoquant y-level as indicated by the blue circle in Figure 3.6. The gap tends to be larger when the data are noisier. However, the gaps can be assured to be zero if we impose homotheticity. In the non-homothetic case, we can always reduce the gap to zero by using fewer rays for estimation, although at the cost of a rougher functional estimate.[8]

3.3.5 Further extensions to the estimation algorithm

As stated above, Algorithm 1 may result in a production function estimate with a gap between the convex isoquant estimates and the S-shape estimates. To address this issue, we develop several extensions to algorithm 2A which allow us to estimate a production function by iterating between the estimations of isoquants and S-shape functions to reduce the size and number of gaps that may exist.

Algorithm 2A is a concise summary of our algorithm. The mathematical details and an extended description is available in Appendix B.2 and is labeled, Algorithm 2B. We use Algorithm

[8] When the gaps are significant, selecting the value for tuning parameters becomes a multi-criteria problem in which we want to minimize both the largest gap and Mean Squared Error (MSE). We do this by setting a threshold on the largest acceptable gap level and picking the tuning parameter value with the smallest MSE. For details of the implementation see Appendix B.2.4.

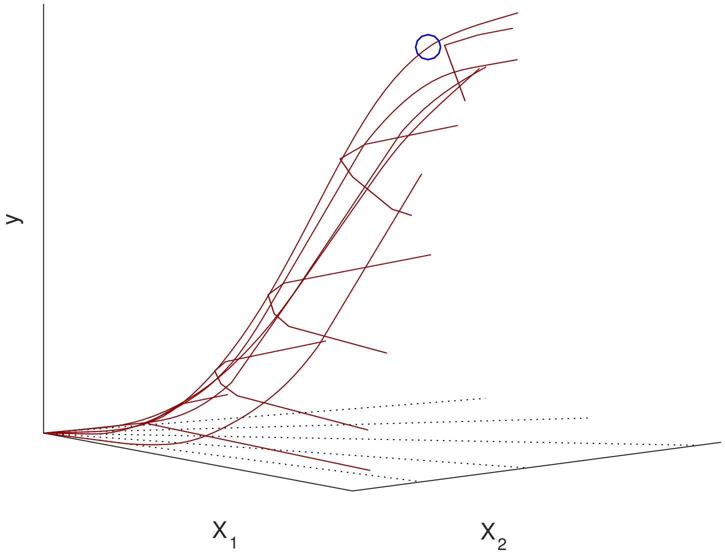


Figure 3.6. Gap between convex isoquant and S-shape estimates

Algorithm 2A Concise summary of the advanced estimation algorithm

```
Data: observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$
procedure
    Initialize the parameters                                                    (B.2.1)
    Estimate each convex isoquant by the CNLS-based method                       (B.2.2)
    Estimate S-shape curve along the rays by the SCKLS-based method              (B.2.3)
    Iterate previous two steps updating the parameters in each iteration until convergence (B.2.4)
return : Estimated function with minimum Mean Squared Errors
```

2B in the following simulation and application sections.

3.4 Simulation study

We use Monte Carlo simulations to evaluate the performance of the proposed estimator with datasets generated by the different data generation process (DGP). We consider both homothetic and non-homothetic functions.

3.4.1 The setup

In our simulation, we compare the performance of the proposed estimator with a Local Linear estimator (LL), which is an unconstrained nonparametric estimation method using kernel weights. We run simulations using the built-in quadratic programming solver, `quadprog`, in `MATLAB`. We define two DGPs with different functional assumptions: homothetic and non-homothetic input isoquants. For each functional assumption, we run experiments varying the sample size and the size of noise. For a testing set drawn from the true DGP, we measure the Root Mean Squared Errors (RMSE) against the true function.

3.4.2 Homothetic DGP

The DGP we use has the following scale function and core function, Olesen and Ruggiero (2014):

$$F(z) = \frac{15}{1 + \exp(-5 \log z)} \tag{3.7}$$

$$g(X_1, X_2; y) = \left(\beta(y) X_1^{\frac{\sigma-1}{\sigma}} + (1 - \beta(y)) X_2^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \tag{3.8}$$

where the elasticity of substitution is $\sigma = 1.51$ and the intensity of the first input, X_1, is $\beta(y) = 0.45$. For the homothetic case, the value of $\beta(y)$ is independent of output level y. We generate samples from

$$y_j = F\left(g(X_{1j}, X_{2j}; y_j^*) \right) + \epsilon_j, \tag{3.9}$$

where y_j^* indicates a true functional value at (X_{1j}, X_{2j}) satisfying

$$y_j^* = F\left(g(X_{1j}, X_{2j}; y_j^*) \right) \tag{3.10}$$

with an additive noise term, ϵ, generated as $\epsilon_j \sim N(0, \sigma_v)$, where σ_v is the standard deviation of the additive noise. We radially generate inputs to the production function, (X_1, X_2), as

$$\boldsymbol{X} = (X_1, X_2) = (\psi \cos \eta, \psi \sin \eta), \tag{3.11}$$

with the modulus, ψ, generated as $\psi \sim unif(0, 2.5)$ and angles, η, generated as $\eta \sim unif(0.05, \frac{\pi}{2} - 0.05)$. Note this DGP specifies that inputs are generated radially and noise is additively contained in the output. This function has homothetic input isoquants because the core function, $g(\cdot)$, is independent of the output level, y.

We consider 9 scenarios varying the training set sample size $(100, 500, 1000)$ and the standard deviations of the noise term, $\sigma_v \in (1.0, 2.0, 3.0)$. We compare our proposed estimator to the LL estimator. To compute the bandwidths for both the LL estimators and the SCKLS estimator for the S-shape part of our algorithm, we use Leave-one-out cross-validation (LOOCV) with the LL estimator. LOOCV is a data-driven bandwidth selection method that has been shown to perform well for unconstrained and constrained kernel estimators, respectively; see Stone (1977) and Yagi et al. (2018).

We use Algorithm 2A to implement our estimator. We specify the number of isoquants and rays as $I = 5$ and $R = 5$, and compute equally spaced percentiles to set the location of the isoquant-level, $\{y^{(i)}\}_{i=1}^{I}$, and rays, $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^{R}$, respectively. We use the average directional CNLS estimates for the isoquant estimation; the details are in Appendix B.2.2.3. We initialize the bandwidth between angles, $\boldsymbol{\omega}$, as $\omega_1 = 0.20$, and increment it by $\Delta\omega = 0.25$. We iterate the procedure 20 times, increasing ω by $\Delta\omega$ in each iteration. After 20 iterations, we select the solution with the smallest sum of squared residuals as our final estimate. We generate 100 training-testing set pairs for each scenario, and draw box plots[9] of RMSE against the true function for both estimators shown in Figure 3.7. The size of the testing set is 1000, and it is randomly drawn from the same distribution as the training set.

We find that the LL estimator performs slightly better than our proposed estimator when the noise is very small, this is likely because our estimator optimizes the fit of the estimated function only on a limited set of grid points. The LL estimator's performance deteriorates as the noise increases because the data-driven nature of LOOCV gives poor estimates of the bandwidth parameters in noisy scenarios. The shape constraints in our iterative estimator make it robust to bandwidth parameters and noisy data. Further, the variance in the performance of our iterative estimator is smaller than that of LL estimator because the shape constraints reduce the estimator's variance. We also find that our iterative estimator has better out-of-sample performance, this is likely because the shape constraints add structures to the estimator, which helps to avoid over-fitting the observations.

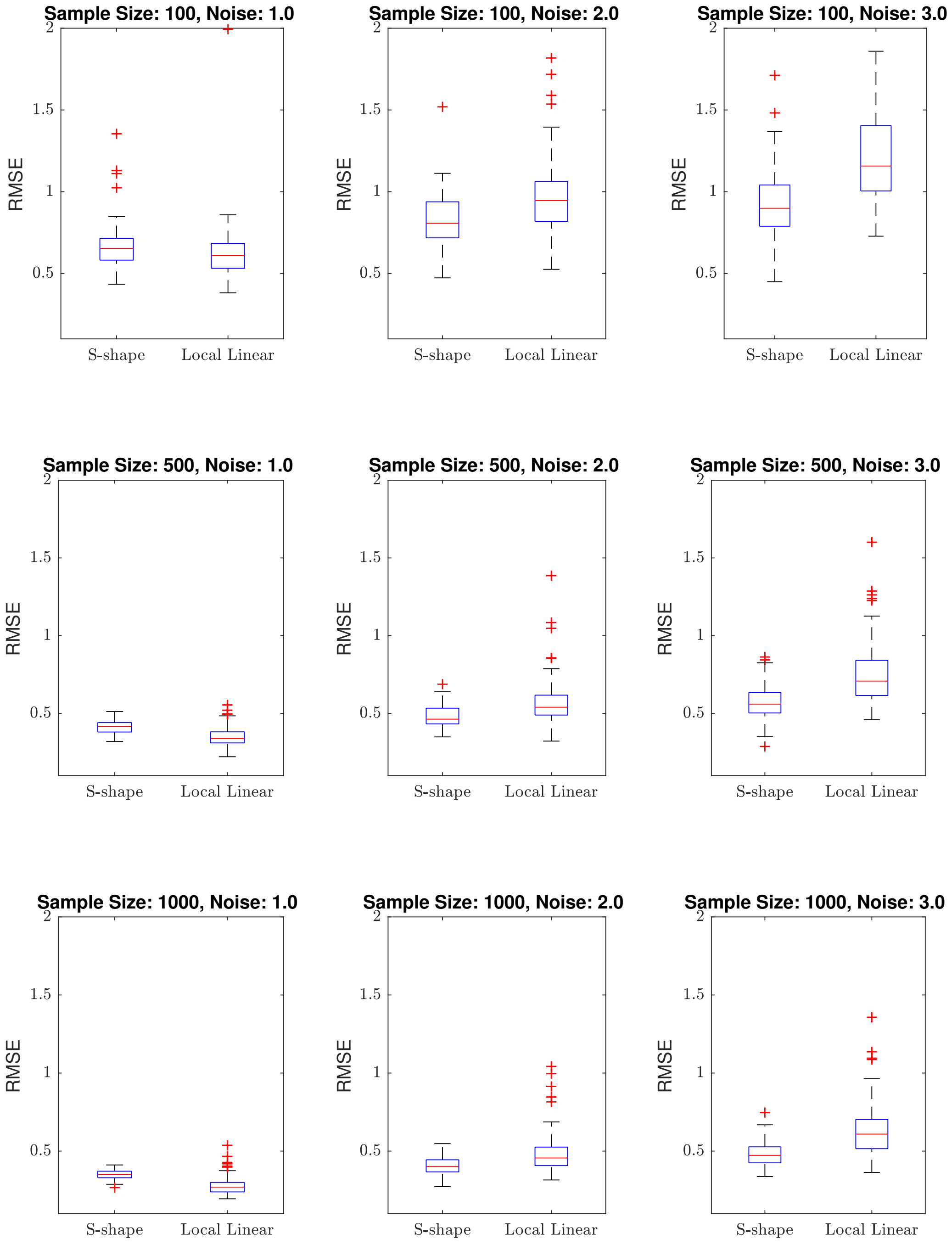
3.4.3 Non-homothetic DGP

We consider the same scale function (3.7) and core function (3.8) as defined in Section 3.4.2. We make the function non-homothetic by redefining the β value as

$$\beta(y) = 0.25 + \frac{y}{15} \times 0.30, \tag{3.12}$$

where $\beta(y) \in [0.25, 0.55]$ depends on the output level $y \in [0, 15]$. We generate the observations by solving equation (3.10) for a given (X_{1j}, X_{2j}). This function is non-homothetic because the core

[9] We define a maximum whisker length of a box plot as $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$, where q_1 and q_3 denote the 25 and 75 percentiles, respectively.


Sample Size: 100, Noise: 1.0

Sample Size: 100, Noise: 2.0

Sample Size: 100, Noise: 3.0

Sample Size: 500, Noise: 1.0

Sample Size: 500, Noise: 2.0

Sample Size: 500, Noise: 3.0

Sample Size: 1000, Noise: 1.0

Sample Size: 1000, Noise: 2.0

Sample Size: 1000, Noise: 3.0

Figure 3.7. Estimation results on the testing sets with a homothetic production function

function $g(\cdot)$ is dependent on an output level y. We run simulations with same settings described in Section 3.4.2, and draw box plots of RMSE values against the true function for each estimator on testing set shown in Figure 3.8.

The results are similar to the homothetic production function. Again, the LL estimator has a larger RMSE variance than our estimator. However, both estimators have larger RMSE variance in the non-homothetic scenarios, particularly in very noisy instances. Our estimator still performs well in terms of RMSE, which indicates its robustness to different assumptions about the production function.

3.5 Application

In this section, we estimate the production function using firm-level industry data from Japan's *Census of Manufactures* provided by METI from 1997 to 2007, when demand for cardboard was relatively constant. Although some researchers have used the same dataset to estimate production functions (Ichimura et al. (2011)), they rely on strong parametric functional assumptions, whereas we relax them and estimate a production function nonparametrically under the RUP law and input convexity. We focus on economic insights related to the cardboard firms' productivity and scale of production.

3.5.1 Census of Manufactures, Japan

The annual *Census of Manufactures* covers all establishments with four or more employees and is conducted by METI under the Japanese Statistics Act. We use establishment-level data with 30 or more employees since the establishment with less than 30 employees do not report capital stock values. We use the same definition of the variables for production functions as Ichimura et al. (2011):

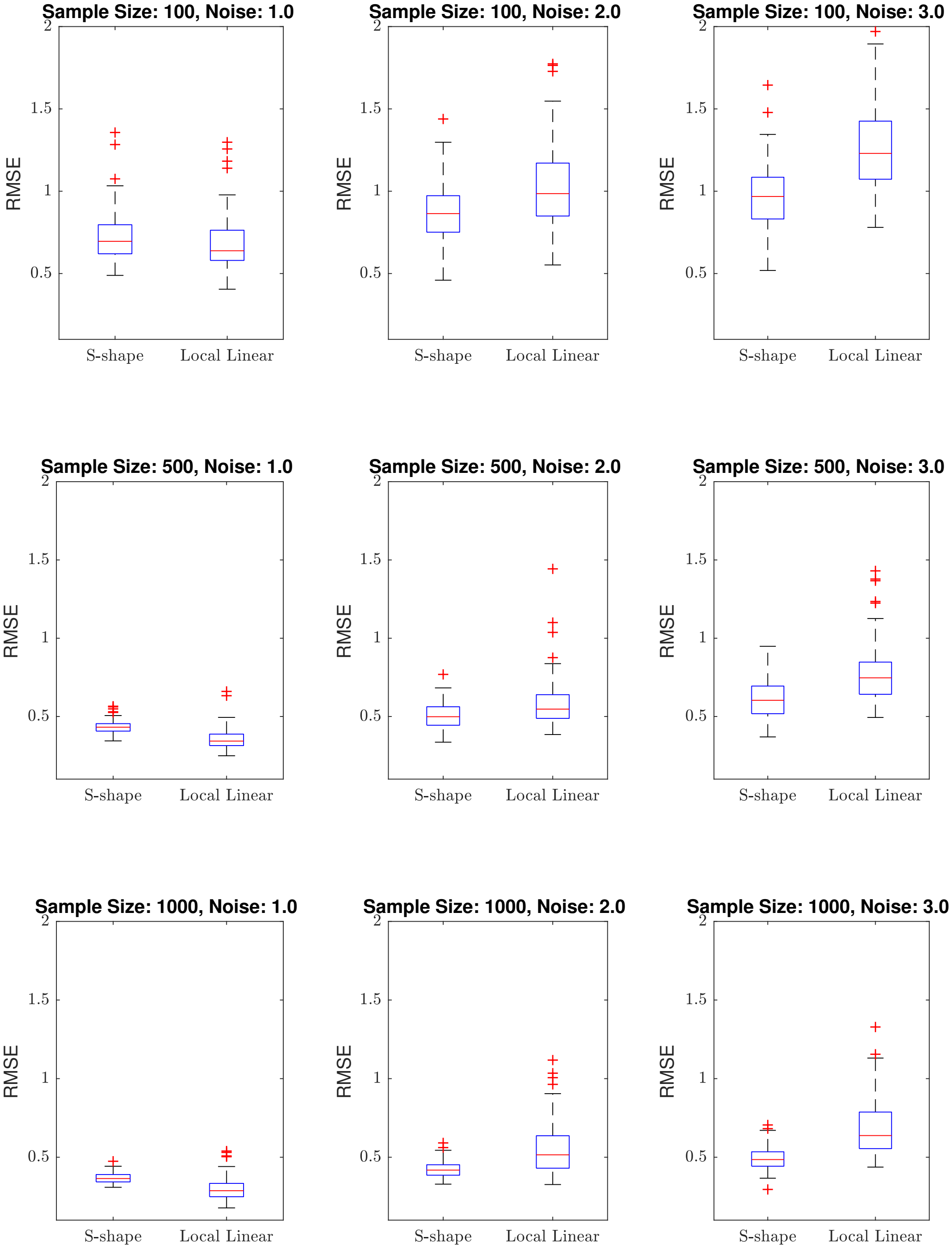


Figure 3.8. Estimation results on the testing sets with a non-homothetic production function

- L = (sum of total regular employees[10] at the end of each month)
- K = (starting amount of tangible assets[11])
- y = (total amount shipped) + (ending inventory of finished and work-in-progress products) - (starting inventory of finished and work-in-progress products) - (cost for intermediate inputs[12])

where L,K and y indicate the labor, capital and value added, respectively, and the production function is modeled as $y = g_0(L, K)$.

We use industry-level deflators obtained from the Japan Industrial Productivity Database (JIP)[13] to convert into year 2000 values. Figure 3.9 shows the price deflator of the cardboard industry and the deflator for Japan's GDP. Note that the price deflator of the cardboard industry is larger than that of GDP after 2003. This finding is consistent with larger firms shrinking their production capacity, which led to higher cardboard prices after 2003, Iguchi (2015).

We convert establishment-level data into firm-level data by summing up the establishment-level data which belong to the same firm. We use firm-level data because expansion decisions are typically made at the firm-level by investing capital, labor, or merging with other firms.

The sample size of the panel data set is $n = 4316$, and there are approximately 400 observations in each year. We normalize each variable by dividing by the standard deviation for data confidentiality. Positive skewness of both the input and output variables implies the existence of many small and a few large firms. Table 3.1 reports the summary statistics.

[10] Regular employees include full-time, part-time, and dispatched workers who work 18 days or more per month.

[11] Tangible assets include machines, buildings, and vehicles.

[12] Intermediate inputs include raw materials, fuel and electricity.

[13] The JIP database is publicly available at *Research Institute of Economy, Trade and Industry* (REITI) (`https://www.rieti.go.jp/en/database/jip.html`)

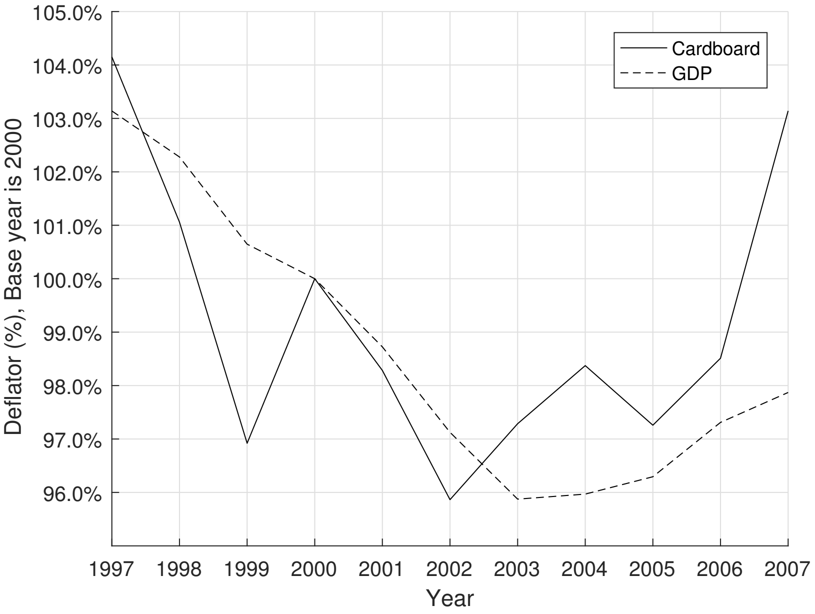


Figure 3.9. Price deflator (Base year = 2000)

Table 3.1. Summary Statistics of the corrugated cardboard industry (1453)

	Labor	Capital	Value added
Mean	0.554	0.283	0.340
Skewness	10.28	11.87	11.86
10-percentile	0.217	0.024	0.059
25-percentile	0.253	0.047	0.093
50-percentile	0.334	0.100	0.158
75-percentile	0.539	0.231	0.298
90-percentile	0.861	0.519	0.567

Figure 3.10, 3.11 and 3.12 show the evolution of each variable across the panel periods by plotting the percentage change of each variable's quartile mean for each year compared with 1997. Here, we compute the quartiles by total amount produced, i.e. firms in the 75%-100% bin have the highest total amount produced, while firms in the lower percentile bin have lower total amount produced. We define total amount produced as:

- (total amount produced) = (total amount shipped) + (ending inventory of finished and work-in-progress products) - (starting inventory of finished and work-in-progress products)

Intuitively, we use the total amount produced as an indicator of a firm's scale size.

The four lines indicate from thinnest to thickest, the 0–25 percentile mean, 25–50 percentile mean, 50–75 percentile mean, and 75–100 percentile mean, respectively. During the time period, firms did not need to adjust their labor levels significantly while most firms reduce their capital levels between 2004 and 2006. We can interpret this as firms in the cardboard industry realized their over-investment in capital and readjusted for more efficient resource use. We observe that the larger firms in our panel dataset expanded value added while reducing their capital levels.

3.5.2 The setup for our application

Before using our iterative algorithm, we specify (1) Number and location of the rays and (2) Number and location (y-levels) of the isoquants. Table 3.1 reports significant skewness of our dataset, i.e. many small firms and only a few large firms. An equally spaced percentile grid will not work well because it may fail to define the rays and isoquant y-levels corresponding to the large firms. Therefore, we use the K-means clustering method to cluster the data into K groups because it is robust to the skewness of our data. However, since K-means clustering requires pre-defining parameter K which is the number of clusters, we use Bayesian Information Criteria (BIC)

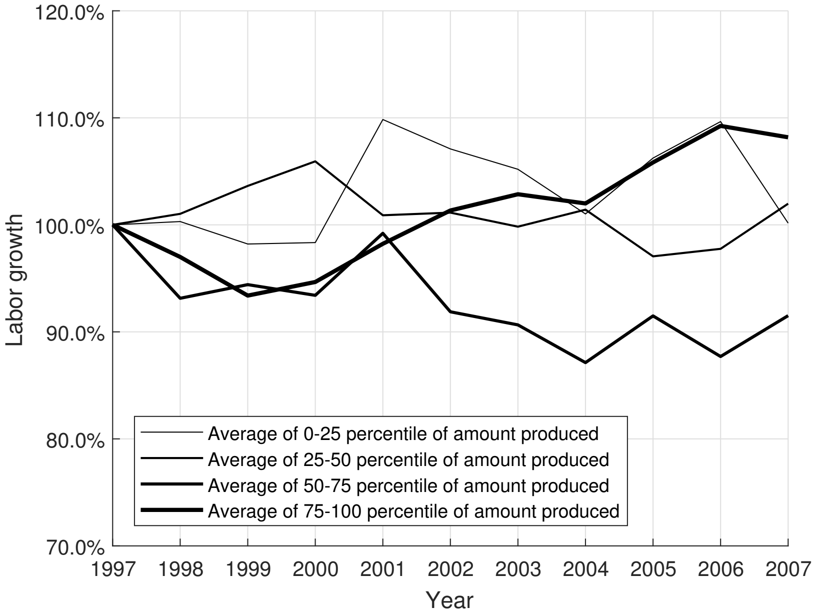


Figure 3.10. Percentage change of quartile mean of labor
(by amount produced, base year = 1997)

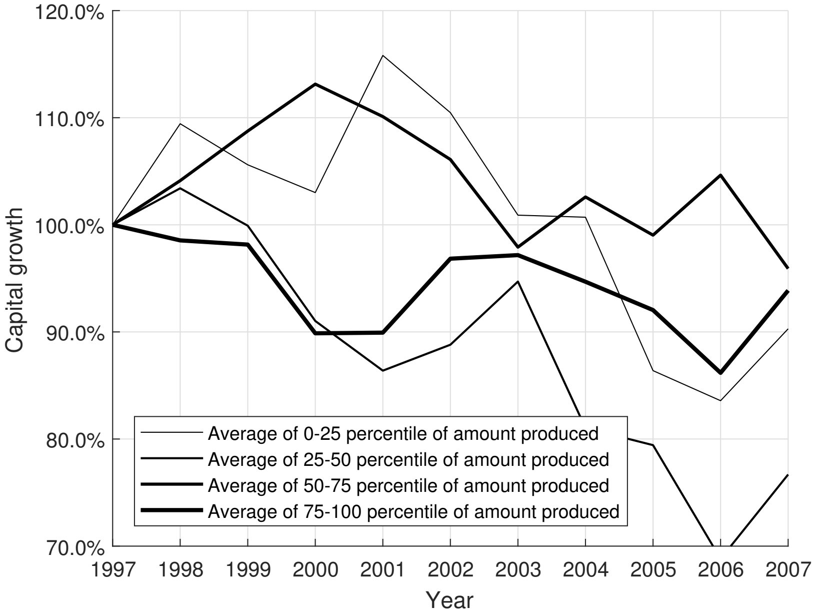


Figure 3.11. Percentage change of quartile mean of capital
(by amount produced, base year = 1997)

to balance the model complexity and explanatory power and avoid over-fitting. We iterate the algorithm 100 times over different K, and find that $K = 12$ provides the lowest BIC value for our dataset. We define the rays and isoquant y-levels as the centroid of each cluster. Figure 3.13 shows the rays and isoquant y-levels defined by K-means clustering. There are many clusters defined for small scale firms and labor intensive firms and there are also a few clusters and associated rays and isoquant y-levels defined for large firms and capital intensive firms.

We initialize the bandwidth between angles as $\omega_1 = 0.20$, and increment it by $\Delta\omega = 0.20$ for each iteration. We iterate the procedure 50 times until ω becomes large enough that the functional estimates are stable between iterations. From the 50 estimates, we select the solution with the smallest sum of squared residuals in our solution set as our final estimate.

3.5.3 Estimated production function and interpretation

Figure 3.14 shows graphs of: (a) the estimated input isoquants, and (b) the estimated S-shape production function on each ray. The black lines indicate the estimates on the centroid of each cluster defined by K-means clustering, and the red points indicate the most productive scale size on each ray from the origin. Figure 3.14 (a) shows that the marginal rate of technical substitution (MRTS) of labor for capital is high when the scale of production is smaller. This indicates that labor is a more important input factor for firms operating at a smaller scale. In contrast, the isoquant becomes flat as the scale of production increases, i.e. the MRTS is low for large firms. These isoquants imply that capital is a more important input factor for larger firms because labor levels need to be increased significantly to offset a small reduction in capital.

Figure 3.14 (b) shows that labor intensive firms have a much smaller most productive scale size than capital intensive firms. This finding coincides with the production economics theory

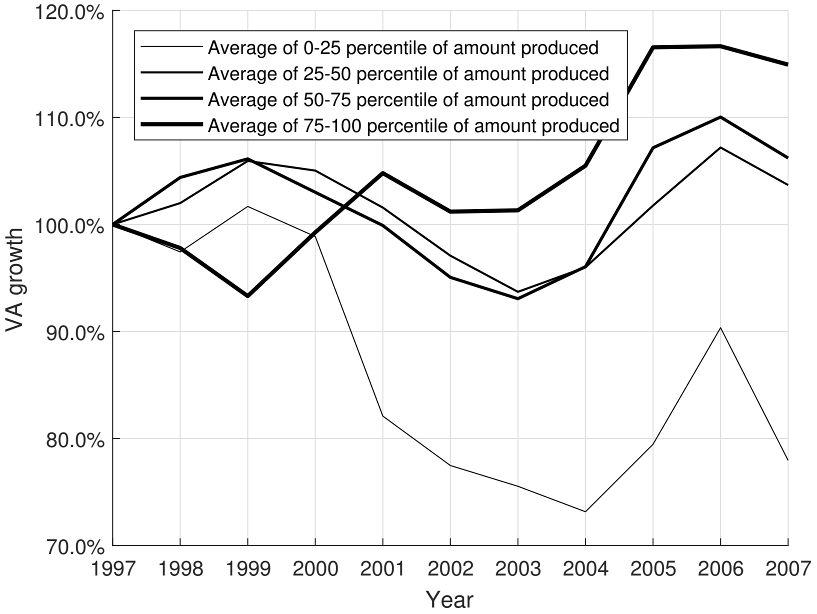


Figure 3.12. Percentage change quartile mean of value added (by amount produced, base year = 1997)

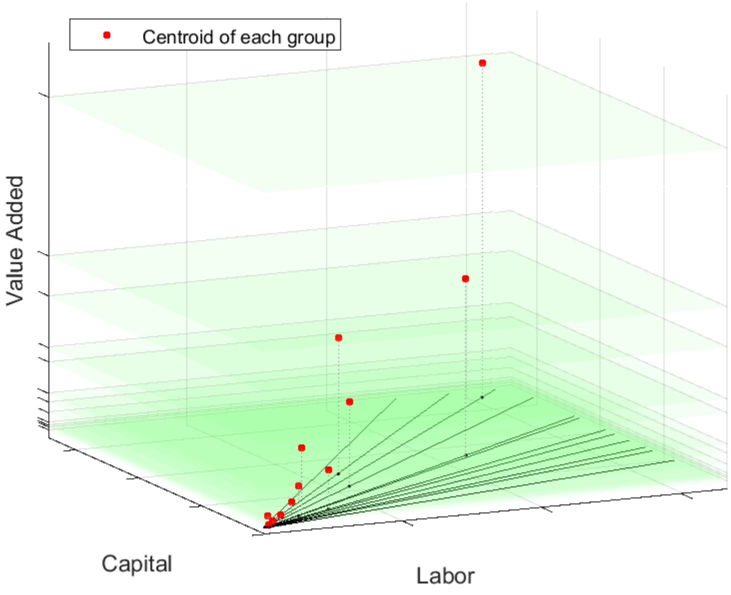
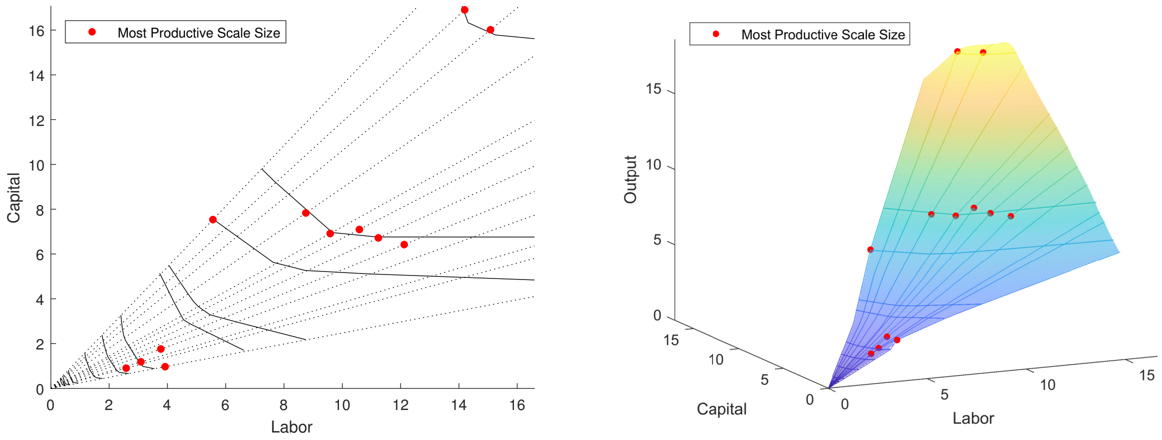


Figure 3.13. Centroid of each group estimated by K-means clustering

stating that firms become more capital intensive as they grow larger by automating processes with capital equipment and using less labor. Note that the most capital intensive ray has a smaller most productive scale size. This is likely a result over-investment in capital. Therefore, firms near this ray could reduce their capital intensity to increase their productivity and scale of operations.



(a) Estimated input isoquants

(b) Estimated production function

Figure 3.14. Estimated results of the corrugated cardboard industry

3.5.4 Analysis on productivity measure

We study how our production function estimator impacts the measures of productivity and productivity variation. Productivity is the ratio of observed output y_{jt} to estimated production function at corresponding input factors, (L_{jt}, K_{jt}). Intuitively, if firms have higher productivity, they can produce larger value added with a given amount of input factors. Productivity can measure the firms' deviation of output (value added) which cannot be explained by the input factors. Syverson (2011) enumerates the primary causes of productivity dispersion as managerial practices, quality of input factors, R&D, learning by doing, product innovation, firms' structure decisions, or other

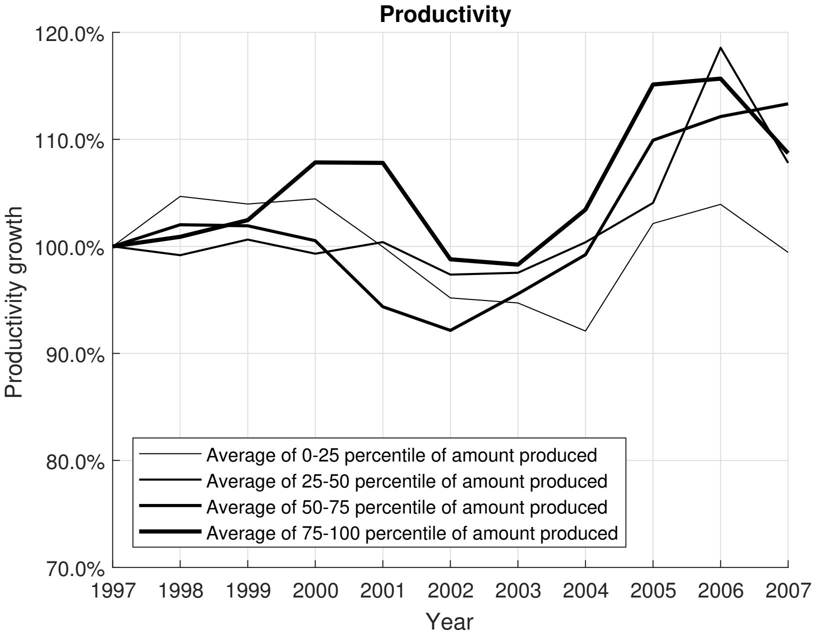


Figure 3.15. Percentage change of quartile mean of productivity
(by amount produced, base year = 1997)

external drivers.

We measure productivity using the concept of total factor productivity (TFP), defined as follows:

$$TFP_{jt} = \frac{y_{jt}}{\hat{g}_0(L_{jt}, K_{jt})} \qquad \forall j = 1, \ldots, n_t \text{ and } \forall t = 1, \ldots, T, \tag{3.13}$$

where n_t is a sample size for each time period t, T denotes the panel periods, and $\hat{g}_0$ is a production function used to aggregate inputs.

First, we investigate how productivity for the cardboard industry is changing over time. Figure 3.15 and 3.16 plot the percentile change of quartile mean of productivity and capital-to-labor input factor ratio for each year compared with 1997, respectively.

Figure 3.15 shows that the medium and large firms have significant productivity growth after 2004, whereas small firms have more stable productivity transition. In contrast, Figure 3.16 de-

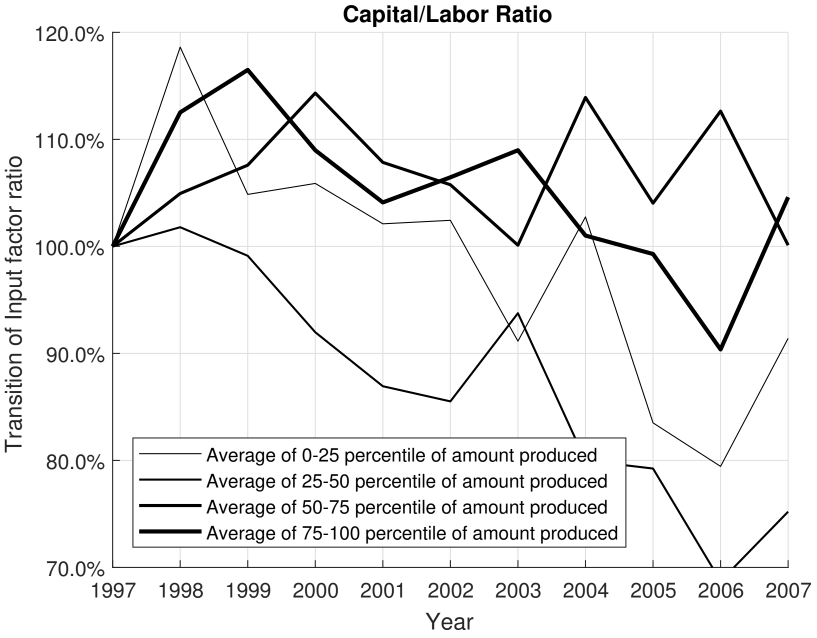


Figure 3.16. Percentage change of quartile mean of input ratio
(by amount produced, base year = 1997)

scribes that firms tend to shrink capital-to-labor ratio after 2004. Thus, we conclude that larger firms improved their productivity significantly by adjusting their input factor ratio for more efficient use of their resource. However, since the productivity of the cardboard industry is heavily dependent on the amount of capital investment, small firms had difficulty to improve their productivity endogenously over the 11 years.

We now turn our attention to the productivity variation observed across firms within the industry. We will use four methods to calculate aggregate inputs. The first two methods are described in Syverson (2004), but we will briefly summarize them here. Aggregate input is estimated by

$$g_0(L_{jt}, K_{jt}) = L_{jt}^{\alpha_L} K_{jt}^{\alpha_K} \tag{3.14}$$

where α_L and α_K are factor elasticities used as weights to aggregate the various inputs. These factor elasticities can be approximated either by industry level cost shares or by individual firm cost shares. Since we have individual firm cost shares in our data set, we calculate both.[14] The third option is to fit the production function using our axiomatic approach. We calculate the estimates, $\hat{g}(\cdot)$ and use these values in Equation 3.13 to calculate TFP.

Table 3.2 summarizes the results of the three methods. Using the industry and firm cost shares results in a 90-10 percentile ratio of 3.97 and 3.56, respectively. This is considerable larger than the the value of 2.68 and 1.91 Syverson (2004) reports as an average across a variety of four digit Standard Industry Classification (SIC) industries in the U.S. economy. However, industries like corrugated cardboard that produce a high homogeneous product are expected to have lower productivity ratio. We find firms in the 90th percentile of the productivity distribution makes almost four times as much output with the same measured inputs as the 10th percentile firm. By relaxing the parametric function form of production function expressed in equation (3.14), using our axiomatic estimator, results in an approximately 25% decrease in productivity variation compared to the industry cost share measure.

Table 3.2. The ratio of the 90th to 10th percentile productivity level for four different methods

	90-10 percentile range
Industry Cost Shares	3.971
Firm Cost Shares	3.559
Axiomatic Estimator	3.167

[14] Because of the various units of measures used for different inputs, the scale of TFP is not easily interpretable. Thus, we normalize each firms TFP by the median TFP for the industry, following Syverson (2004).

3.6 Conclusion

This chapter develops an approach to estimate a general production function imposing economic axioms, both the RUP law and the input convexity. The axioms can be stated as shape constraints and the proposed estimator is implemented as a non-parametric shape constrained regression. This approach allows considerable more flexibility then widely used parametric methods.

We use these developments to analyze a panel dataset of Japan's cardboard industry from 1997 to 2007. We observe a capacity contraction after 2004 across most of the larger firms in the industry. The contraction's timing corresponds to an increase in the price index for cardboard productions, indicating increasing market power of firms in the industry. We estimate the production function and compute the most productive scale size and the productivity of each firm. We find most productive scale size is significantly influence by the capital-labor ratio of the firm. In particular firms with higher capital-to-labor ratios have a larger most productive scale size than firms with lower capital-to-labor ratios. Productivity variation is small relative to other industries and countries. However, using an axiomatic estimator that accounts for productivity variations due to scale and input mix reduces unexplained productivity variation by approximately 25%.

In next chapter, we plan to extend our analysis to other industries in Japan which have roughly homogeneous outputs such as bread, coffee, concrete, plywood, and sugar. Census of Manufactures data are self reported by firms and are notoriously noisy. We will test the data in each industry to see if simpler parametric models or alternative non-parametric shape constrained models are sufficient to capture the main characteristics for the data. We will study the patterns across industries to identify which factors consistently influencing productivity.

As managers strategically plan the expansion of their firm, estimates of the most productive scale size, the trade-offs between manual and automated operations, and the potential outputs gains to expansion provide critical insights to the benefit-cost analysis. The proposed axiomatic approach imposes a minimum set of axioms that still allows for the standard interpretation of the production function allowing managers to be better informed when taking critical planning decisions for the firm.

4. SHAPE CONSTRAINED NONPARAMETRIC IV ESTIMATOR AND ITS APPLICATION TO PRODUCTION FUNCTIONS

4.1 Introduction

Endogeneity is believed to be a common problem when estimating production function since firms' managers determine input levels while perceiving a part of their firm specific productivity level which is modeled as a part of the residual. Specifically, firms change variable input (labor) levels due to their productivity level, which makes typical regression estimates biased and inconsistent. For instance, when firms observe the higher productivity level, then they may higher more labors in this year to prepare for a busy period in the near future. Marschak and Andrews (1944) were among the first to point out endogeneity of observed input demands as rational production managers adjust their input use for technical inefficiency. Several solutions to this endogeneity problem have been suggested. Marschak and Andrews (1944) propose to estimate a system of demand equations, and hence the endogeneity of input variables is often referred to as the simultaneity problem. Other standard econometric approaches to address endogeneity of this type include the use of instruments or panel data (e.g., Mundlak (1961); Mundlak and Hoch (1965)). Zellner et al. (1966) were the first to discuss in detail the timing of the input consumption decisions relative to when the productivity shocks are observed. The empirical industrial organization literature has focused on this type of solution. There is other solution for endogeneity called the control function approach which uses some proxy variables to explain a part of residuals correlated with input. Olley and Pakes (1996) propose to avoid endogeneity of inputs by using investment as a proxy for productivity. More recent studies in that stream include Levinsohn and Petrin (2003) and Ackerberg et al. (2015) who explore the use of other proxy variables for productivity such as material inputs. Although these approaches to deal with endogeneity play a key role in theory and application of production economics, their models are dependent on a parametric functional framework.

Recently, nonparametric instrumental variables (IV) approach is proposed to deal with endogeneity with a flexible functional form. Newey et al. (1999) uses 2–stage nonparametric estimation with a control function approach to explain a part of residuals correlating with endogenous variables. Newey and Powell (2003) consider a more general simultaneous equation models with series–based estimator. Kernel–based approach with Tikhonov regularization is proposed by Hall and Horowitz (2005) and Darolles et al. (2011). Florens et al. (2018) proposed kernel–based approach with Landweber–Fridman regularization techniques which require iterations, and they estimate the marginal effect of instrumental variables.

In this paper, we propose a shape constrained nonparametric IV estimator which imposes a set of shape constraints on a nonparametric IV approach. We apply the Landweber–Fridman regularization to the Shape Constrained Kernel–weighted Least Squares (SCKLS) estimator developed by Yagi et al. (2018). Furthermore, we also consider more complicated shape constraints proposed by microeconomic theory by applying iterative S–shape algorithm proposed by Yagi et al. (2018). We aim to improve the finite sample performance and the economic interpretability of estimated results by imposing correctly specified shape constraints while avoiding the bias from endogeneity issues.

We estimate production functions for the following Japanese manufacturing industries which produce highly homogeneous products: sugar, bread, coffee, plywood, cardboard, ready-mixed concrete and concrete products. We use full–time labor headcount as IV for the labor input since managers are not able to adjust full–time labors easily in Japan and is highly correlated with the overall labor consumed by the firm. Instead, firms try to adjust the labor input through part–time labors, and thus, we control the endogeneity in the labor input by using full–time labors as IV. The estimation results provide a description of supply–side of Japanese manufacturing industry as we report industry–level aggregated productivity and the most productive scale size. We find the model specification changes the productivity estimates significantly and may lead to different interpretations and economic insights. Specifically, restrictive parametric models such as Cobb–Douglas two–stage least squares (2SLS) are likely to be suffered from the bias due to the misspecification. Meanwhile, models without IV will be biased in case that endogeneity exists in the data set.

The chapter is structured as follows. The assumption and model framework is described in Section 4.2. Monte Carlo simulation results under different Data Generation Process (DGP) reflecting practical scenarios are shown in Section 4.3. Section 4.4 applies the proposed estimator to analyze the productivity for the Japanese manufacturing industries. Finally, Section 4.5 concludes and suggests potential further research directions.

4.2 Model and methodology

4.2.1 Model

Consider the following production function model

$$y = g_0\left(\boldsymbol{x}\right) + \epsilon, \tag{4.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{d_1}$ is d_1–dimensional input vector, y is an output scalar, $g_0(\cdot)$ is a production function to be estimated and ϵ is unobserved residuals. Under endogeneity, the conventional exogenous assumption on independent variables, $E[\epsilon|\boldsymbol{x}] = 0$, does not hold because of the correlation between residuals and independent variables. In many of existing researches, such as Olley and Pakes (1996) and Levinsohn and Petrin (2003), the model assumes strong parametric functional forms on a production function g_0. Here we aim to relax these restrictive assumptions by estimating functions nonparametrically with general structures which can be stated as shape constraints such as concavity, S–shape or monotonicity, which help to maintain the interpretability of estimates. Given instrumental variables, $\boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^{d_2}$ is d_2–dimensional vector and satisfies the exclusion restriction $E[\epsilon|\boldsymbol{z}] = 0$, we can take a conditional expectation of both sides of Equation (4.1) on $\boldsymbol{z}$,

$$E[y|\boldsymbol{z}] = E[g_0(\boldsymbol{x})|\boldsymbol{z}]. \tag{4.2}$$

The conditional expectation of Equation (4.2) yields the following integral equation

$$E[y|\boldsymbol{z}] = \int_{-\infty}^{\infty} g_0(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{z})\, d\boldsymbol{x}. \tag{4.3}$$

where $f(\boldsymbol{x}|\boldsymbol{z})$ is a probability density function of $\boldsymbol{x}$ conditional on $\boldsymbol{z}$. We also assume that the random vector $\boldsymbol{x}$, y and $\boldsymbol{z}$ are distributed based on their density function $f_{\boldsymbol{x}}$, f_y and $f_{\boldsymbol{z}}$, and define the Hilbert space of square integrable functions depend on each random vectors by $L^2(\boldsymbol{x})$, $L^2(y)$ and $L^2(\boldsymbol{z})$.

Now we focus on estimating a production function $g_0(\cdot)$. The mapping from $g_0(\boldsymbol{x})$ to $E[y|\boldsymbol{z}]$ is continuous if $f(\boldsymbol{x}|\boldsymbol{z})$ is bounded; however, an inverse mapping is not continuous, which means small changes in $E[y|\boldsymbol{z}]$ do not produce small changes in $g_0(\cdot)$. This is referred to as the ill–posed inverse problem and requires regularization which modifies the inverse mapping to remove the discontinuity. The amount of modification is controlled by a positive constant called regularization parameter. Intuitively speaking, with a large regularization parameter, the revised inverse mapping becomes more stable but a bad approximation for the inverse mapping. In contrast, with a small regularization parameter, the revised inverse mapping becomes a good approximation for the inverse mapping but unstable. Thus, it is also important to select an optimal regularization parameter which minimizes the trade–off between approximation and stability. See Horowitz (2014) for more details on the ill–posed problems and solutions.

One of the regularization methods widely used in a nonparametric IV approach is called Tikhonov regularization. Here, equation (4.2) can be written as

$$\phi = Tg_0, \tag{4.4}$$

where $\phi = E[y|\boldsymbol{z}]$ and T is an operator defined by $Tg_0 = E[g_0\left(\boldsymbol{x}\right)|\boldsymbol{z}]$. More specifically,

- $T : L^2(\boldsymbol{x}) \to L^2(\boldsymbol{z})$ such that $g \to Tg = E[g\left(\boldsymbol{x}\right)|\boldsymbol{z}]$
- $T^* : L^2(\boldsymbol{z}) \to L^2(\boldsymbol{x})$ such that $h \to T^*h = E[h\left(\boldsymbol{z}\right)|\boldsymbol{x}]$

where T^* is called adjoint operator which satisfies a following condition:

$$\langle Tg(\boldsymbol{z}), h(\boldsymbol{z})\rangle = \langle g(\boldsymbol{x}), T^*h(\boldsymbol{x})\rangle$$

Then the Tikhonov regularization is calculated by solving the following optimization problem:

$$\min_{g_0} \left\|Tg_0 - \phi\right\|^2 + \delta \left\|g_0\right\|^2 \tag{4.5}$$

where $\delta > 0$ is a regularization parameter. The solution of this problem is given by

$$g_0^\delta = \left(\delta I + T^* T\right)^{-1} T^* r \tag{4.6}$$

This is analogous to a ridge regression method widely used for solving multicollinearity problems in the field of regression analysis. Darolles et al. (2011) applied this regularization for solving the nonparametric IV approach with a kernel-weighted estimator. However, this regularization method requires to compute the inversion of $n \times n$ matrix, the problems becomes difficult to solve within a feasible time when sample size n is large. This makes Tikhonov regularization difficult to apply to a large-scale data set such as estimating industry specific production functions from manufacturing census data.

Alternatively, the Landweber–Fridman iterative regularization method is computationally easier than Tikhonov regularization. The Landweber–Fridman iterative method is implemented by solving the following minimization problem:

$$\min_{g_0} f\left(g_0\right) = \frac{1}{2}\left\|T g_0 - \phi\right\|^2 . \tag{4.7}$$

The iterative algorithm is given by updating g_0 using the gradient of the objective function $\nabla f\left(g_0\right)$ such that

$$\begin{aligned} g_{0,k+1} &= g_{0,k} - c \nabla f\left(g_0\right) \\ &= g_{0,k} - c T^* \left(T g_{0,k} - \phi\right) \end{aligned} \tag{4.8}$$

where $c < 1$ is a positive constant called the relaxation parameter which controls the step size of updates. This can be seen as an special case of gradient descent algorithm widely used in the field of mathematical optimization. For convergence of this iterative algorithm, we require following condition on the parameter c:

$$
\begin{aligned}
\left\|g_{0,k+1} - g_{0,k}\right\|^2 &= c \left\|T^* \left(T g_{0,k} - \phi\right)\right\|^2 \\
&= c \left\|T^* \left(T g_{0,k} - T g\right)\right\|^2 \\
&\leq c \left\|T^* T\right\|^2 \left\|g_{0,k} - g\right\|^2 \quad < \left\|g_{0,k} - g\right\|^2 .
\end{aligned}
$$

This requires to choose a positive constant $c > 0$ such that $c \left\|T^* T\right\|^2 < 1$. This algorithm stops when the deviation $\left\|T g_0 - \phi\right\|^2$ approaches to the noise level of the estimator.

Florens et al. (2018) applied Landweber–Fridman iterative method to a nonparametric IV estimation where their main focus is to evaluate the marginal effects of IV regression. More specifically, they estimate the unknown operator and function with kernel regression such as Local Linear (LL) estimator, and iterate Landweber–Fridman algorithm until some convergent conditions are satisfied. Here, we impose the shape constraints on the nonparametric estimation to improve the finite sample performance and the interpretability of estimation results.

4.2.2 Unconstrained estimator

We can replace the unknown functions and operators in Equation (4.8) with a local–weighted kernel regression estimator. Here we summarize the algorithm step–by–step to estimate the unconstrained nonparametric IV estimator.

For each step, we implement a local–weighted kernel regression such as Local Linear estimator. Bandwidth for each kernel estimation can be selected by data–driven methods such as leave–one–out cross validation. We suggested to use a similar stopping criteria to Florens et al. (2018) that the algorithm will stop when the deviation across iterations becomes small. Specifically, we consider

Algorithm 2B Nonparametric IV estimator

```
Data: observations $\{\boldsymbol{X}_j, \boldsymbol{Z}_j, y_j\}_{j=1}^{n}$
procedure
Initialization:
    $k \leftarrow 0$
    Estimate a starting points $g_{0,0}\left(\boldsymbol{X}_j\right) = E[y_j|\boldsymbol{X_j}]$
Iteration:
    while Stopping criteria are not satisfied do
        Given an estimate from the previous iteration $\hat{g}_{0,k}$,
        estimate $T\hat{g}_{0,k} - \phi = E[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]$
        Given an estimate of the previous step,
        estimate $T^*\left(\hat{T}\hat{g}_{0,k} - \hat{\phi}\right) = E[\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]|\boldsymbol{X}_j]$
        Update a production function with
        $\hat{g}_{0,k+1} = \hat{g}_{0,k} - c\hat{T}^*\left(\hat{T}\hat{g}_{0,k} - \hat{\phi}\right) = \hat{g}_{0,k}(\boldsymbol{X}_j) - c\hat{E}[\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]|\boldsymbol{X}_j]$
        $k \leftarrow k + 1$
    end
return : A production function of the last iteration $\hat{g}_{0,k}$
```

the following normalized deviation:

$$\left\|\frac{\hat{T}\hat{g}_{0,k} - \hat{\phi}}{\hat{\phi}}\right\|^2 = \left\|\frac{\hat{E}[g_{0,k}|\boldsymbol{Z}_j] - \hat{E}[y_j|\boldsymbol{Z}_j]}{\hat{E}[y_j|\boldsymbol{Z}_j]}\right\|^2, \tag{4.9}$$

and we propose to stop the iteration when the difference of the normalized deviation across iterations become smaller than some threshold $\delta > 0$.

$$\left|\left\|\frac{\hat{T}\hat{g}_{0,k-1} - \hat{\phi}}{\hat{\phi}}\right\|^2 - \left\|\frac{\hat{T}\hat{g}_{0,k} - \hat{\phi}}{\hat{\phi}}\right\|^2\right| < \delta. \tag{4.10}$$

4.2.3 Constrained estimator with SCKLS

In this section, we propose to extend the unconstrained IV estimator with the SCKLS estimator developed by Yagi et al. (2018). The SCKLS estimator imposes shape constraints on a production function by restricting the function's partial derivatives to impose economic axioms such as supermodularity, convexity, monotonicity, and quasi-convexity. In this implementation, we focus on a global concavity and monotonic increasing constraints allowing us to impose a basic characteristics of a production function for matured industry, decreasing returns to scale (DRS).

Specifically, we propose to combine Step 9 and 10 in Algorithm 2B, and impose the shape constraints on the updated production function $\hat{g}_{0,k+1}$ when we estimate the conditional expectation $E[\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]|\boldsymbol{X}_j]$.

First we introduce a set of m points, $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, for evaluating constraints, which we call evaluation points. Under monotonicity and concavity constraints, we define the production function estimates at k–th iteration as followed:

$$\hat{g}_{0,k}(\boldsymbol{x}; \hat{\boldsymbol{a}}_k, \hat{\boldsymbol{b}}_k) = \min_{i \in \{1,\ldots,m\}} \left\{ \hat{a}_{k,i} + (\boldsymbol{x} - \boldsymbol{x}_i)' \hat{\boldsymbol{b}}_{k,i} \right\} \tag{4.11}$$

where $\hat{a}_{k,i}$ is a functional estimate and $\hat{\boldsymbol{b}}_{k,i}$ is a slope estimate at the evaluation point $\boldsymbol{x}_i$. This implies the estimated production function is a piece–wise linear function since the functional estimate is constructed by taking the minimum of linear interpolations between the evaluation points.

Next, we define the estimate of the conditional expectation $E[\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]|\boldsymbol{X}_j]$ in the spirit of Local Linear kernel estimation. Specifically, we define $\tilde{a}_i$ as a function value and $\tilde{\boldsymbol{b}}_i$ as a slope of $E[\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]|\boldsymbol{x}_i]$ at the evaluation point $\boldsymbol{x}_i$, which describes the correction term for the endogeneity. Then the SCKLS estimator for Step 9 in Algorithm 2B is formulated as follows:

$$
\begin{aligned}
&\min_{\tilde{a}_i, \tilde{\boldsymbol{b}}_i} && \sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j] - \tilde{a}_i - (\boldsymbol{X}_j - \boldsymbol{x}_i)'\tilde{\boldsymbol{b}}_i)^2 K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}_i}{\boldsymbol{h}}\right) \\
&\text{subject to} && (\hat{a}_{k,i} + c\tilde{a}_i) - (\hat{a}_{k,l} + c\tilde{a}_l) \geq \left(\hat{\boldsymbol{b}}_{k,i} + c\tilde{\boldsymbol{b}}_i\right)'(\boldsymbol{x}_i - \boldsymbol{x}_l), \quad i, l = 1, \ldots, m \\
& && \hat{\boldsymbol{b}}_{k,i} + c\tilde{\boldsymbol{b}}_i \geq 0, \qquad\qquad i = 1, \ldots, m.
\end{aligned}
\tag{4.12}
$$

where $K\left(\frac{\boldsymbol{X}_j - \boldsymbol{x}}{\boldsymbol{h}}\right)$ denotes a product kernel, and $\boldsymbol{h}$ is a vector of bandwidths (see Racine and Li (2004) for more detail). Note that $\hat{E}[\hat{g}_{0,k}(\boldsymbol{X}_j) - y_j|\boldsymbol{Z}_j]$ is given by Step 8 in Algorithm 2B, and $\hat{a}_{i,k}$ and $\hat{\boldsymbol{b}}_{k,i}$ are the values computed in the previous iteration. The first set of constraints in (4.12) imposes concavity and the second set of constraints imposes monotonic increasing on a estimate of the updated production function $\hat{g}_{0,k+1}$ at each evaluation point $\boldsymbol{x}_i$. For more details about these constraints, see Kuosmanen (2008). After solving the problem (4.12), now we can update the production function as:

$$
\begin{aligned}
\hat{g}_{0,k+1}(\boldsymbol{x}; \hat{\boldsymbol{a}}_{k+1}, \hat{\boldsymbol{b}}_{k+1}) &= \min_{i \in \{1,\ldots,m\}} \left\{\hat{a}_{k+1,i} + (\boldsymbol{x} - \boldsymbol{x}_i)'\hat{\boldsymbol{b}}_{k+1,i}\right\} \\
&= \min_{i \in \{1,\ldots,m\}} \left\{ \left(\hat{a}_{k,i} + c\hat{\tilde{a}}_i\right) + (\boldsymbol{x} - \boldsymbol{x}_i)'\left(\hat{\boldsymbol{b}}_{k,i} + c\hat{\tilde{\boldsymbol{b}}}_i\right) \right\}
\end{aligned}
\tag{4.13}
$$

where $\hat{\tilde{a}}_i$ and $\hat{\tilde{\boldsymbol{b}}}_i$ are the functional and slope estimates obtained by the optimization problem (4.12). This corresponds to Step 10 in Algorithm 2B.

We propose to implement the SCKLS estimator in Step 9 in Algorithm 2B either (a) at last iteration or (b) at each iteration. If we impose shape constraints only at the last iteration, we continue to use the unconstrained estimator before the last iteration, and implement the SCKLS estimator only once at the last iteration. If we choose to implement shape constraints at each iteration, we solve the SCKLS estimator defined in Equation (4.12) at each iteration in Step 9 in Algorithm 2B, and continue to update the estimates of a production function. In case that we impose shape constraints at each iteration, we expect that the estimator will converge faster than if the shape constraint is only imposed at the last iteration since shape constraints help to improve the finite sample performance. We will compare the performance of these two estimators in Section 4.3 through Monte Carlo simulations.

4.2.4 Constrained estimator with S–shape estimator

In this section, we extend the unconstrained IV estimator by using the iterative S–shape estimation proposed by Yagi et al. (2018). The algorithm aims to estimate the production function satisfying both the RUP law and input convexity, which is a minimal set of economic axioms for a production function that are unlikely to be violated. The RUP law states that along any expansion path, the production function should first have increasing returns to scale followed by decreasing returns to scale, Frisch (1964). These shape restrictions assure we have well-defined marginal products and most productive scale sizes, and thus, will improve the interpretability of the estimation results in addition to the improvement of the finite sample performance.

Unlike the extension with the SCKLS estimator, we cannot combine the iterative S–shape algorithm with the estimation in Step 9 in Algorithm 2B since the S–shape function is estimated only on the expansion rays from the origin. Thus, we propose to apply the iterative S–shape estimator to the unconstrained IV estimator in Section 4.2.2. Intuitively, after we obtained the unconstrained estimator $\hat{g}_0(\boldsymbol{X}_j)$, we estimation the S–shape function with $\{\boldsymbol{X}_j, \hat{g}_0(\boldsymbol{X}_j)\}_{j=1}^{n}$ by using the iterative S–shape estimation. In other words, with the iterative S–shape estimation, we aim to find the closest shape constrained estimator to the unconstrained IV estimator obtained by

Algorithm 2B.

4.3 Simulation study

In this section, we use Monte Carlo simulations to evaluate the finite sample performance and robustness of the proposed estimators. Data sets are randomly generated using different production functions based on the state of industry: emerging, growing and matured.These states of the industry correspond to different shape restrictions.

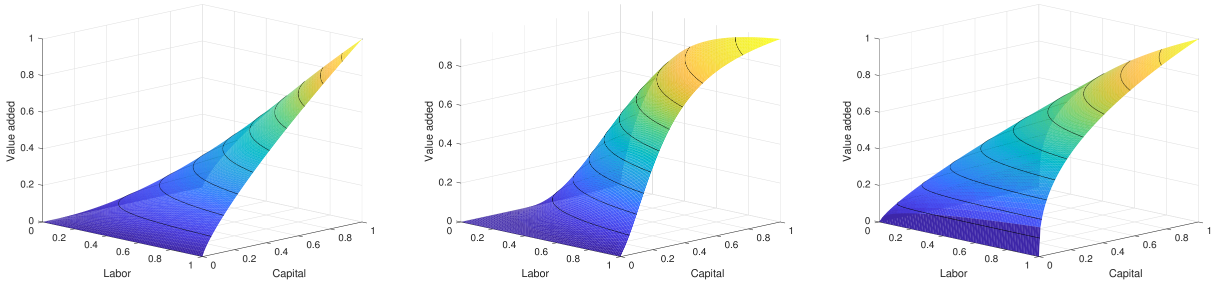
4.3.1 The setup and DGP

We compare the performance of shape constrained nonparametric IV estimator with non–IV approach (LL, SCKLS and Iterative S–shape estimator) and unconstrained IV estimator with LL estimator. We use `MATLAB` to implement these estimators and solve the quadratic programming problems with the build–in quadratic solver, `quadprog`.

We use the Gaussian kernel and leave–one–out cross–validation (LOOCV) for bandwidth selection. We apply LOOCV to all of local–weighted kernel estimators. For the SCKLS estimator, we use 100 evaluation points, which implies a 10 by 10 grid. For the iterative S–shape, we use five rays and five isoquant y–levels to approximate S–shape function with an initial bandwidth between rays of 0.25 and the bandwidth increases by 0.25 at each iteration. See Yagi et al. (2018) for more details. For the Landweber–Fridman regularization, we use $c = 0.5$ and a stopping criteria defined in the equation (4.10) with $\delta = 0.01$.[1]

Since our focus is on production function estimation, we define the DGP based on the different states of the industry. First, we run our experiments for an emerging industry, which implies that the most of firms experience increasing returns to scale due to the relative youth of the firms in the

[1] Florens et al. (2018) shows that $c = 0.5$ empirically provides a good balance between precision and computational time.



(a) Emerging industry

(b) Growing industry

(c) Matured industry

Figure 4.1. Production function used in the simulations

industry. Second, we use a production function for a growing industry, which means smaller firms have increasing returns to scale while larger firms have decreasing returns to scale. Intuitively, the shape of production function is S–shaped. Finally, we use a production function for a mature industry where most of firms face decreasing returns to scale. This indicates that firms will have decreasing marginal products as they get larger due to the matured nature of the industry and challenges related to increasing span of control. Figure 4.1 provides the shape of a true production function for each industry we consider in our experiments.

As shown in Figure 4.1, we focus on the production function with two–input: Labor and Capital, and denote them as X_L and X_K respectively. It is common that labor is correlated with the unobserved productivity because firms can change the number of employees flexibly in the short–term based on their productivity–level. Since we as an analyst cannot observe this productivity–level from the census data, unobserved productivity will be included as part of the unobserved residual, and thus, we have an endogeneity issue, $E[\epsilon|X_L] \neq 0$. We also assume that there is a single instrumental variable which is full–time workers, Z. We set our output measure as value–added

y which is expressed as a function of labor and capital,

$$y = g_0(X_L, X_K) + \epsilon.$$

We consider six different scenarios with different training data sizes $\{100, 500, 1000\}$ and correlations between labor and the IV $\{0.3, 0.6\}$. The correlation between labor and the IV describes the strength of the IV. If the correlation is small, then the IV is weak so that the IV does not have power to explain the variation of endogenous variable uncorrelated with the residuals. We compute the root mean squared errors (RMSE) on 1,000 testing data randomly drawn from the same distribution as the training data.

Inputs, IVs, and residuals are randomly drawn from multivariate normal distribution. Specifically,

$$\begin{pmatrix} X_{Lj} \\ X_{Kj} \\ Z_j \\ \epsilon_j \end{pmatrix} \sim N(\boldsymbol{M}, \boldsymbol{\Sigma}) \qquad \text{for } j = 1, \ldots, n$$

$$\text{where } \boldsymbol{M} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_L^2 & \rho_{LK}\sigma_L\sigma_K & \rho_{LZ}\sigma_L\sigma_Z & \rho_{L\epsilon}\sigma_L\sigma_\epsilon \\ \rho_{LK}\sigma_L\sigma_K & \sigma_K^2 & \rho_{KZ}\sigma_K\sigma_Z & 0 \\ \rho_{LZ}\sigma_L\sigma_Z & \rho_{KZ}\sigma_K\sigma_Z & \sigma_Z^2 & 0 \\ \rho_{L\epsilon}\sigma_L\sigma_\epsilon & 0 & 0 & \sigma_\epsilon^2 \end{pmatrix},$$

$$\begin{pmatrix} \sigma_L \\ \sigma_K \\ \sigma_Z \\ \sigma_\epsilon \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.20 \end{pmatrix}, \text{ and } \begin{pmatrix} \rho_{LK} \\ \rho_{LZ} \\ \rho_{KZ} \\ \rho_{L\epsilon} \end{pmatrix} = \begin{pmatrix} 0.30 \\ \{0.3, 0.6\} \\ 0.10 \\ 0.75 \end{pmatrix}.$$

4.3.2 Emerging industry

In this scenario, we assume that the true production function is a Cobb–Douglas function with increasing returns to scale.

$$g_0(X_L, X_K) = X_L^{0.8} X_K^{0.8}. \tag{4.14}$$

We recognize the selection of coefficients of Cobb–Douglas function may be extremely large for practical experiments. However, these parameter values demonstrate the estimator's are robust to different states of the industry.

Figure 4.2 shows the estimators performance over 50 simulations using box–plots[2] and the RMSE metric. For the various DGPs, the underlined estimators indicate the estimators with cor-

[2] We define a maximum whisker length of a box plot as $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$, where q_1 and q_3 denote the 25 and 75 percentiles, respectively

rectly specified models. For emerging industries, only LL with IV and S-shape with IV estimators are correctly specified.

Figure 4.2 shows that the performance of IV estimators is significantly improved if we have strong IVs, which means high correlation between the endogenous variables and the IVs ρ_{LZ}. Furthermore, we also observe that the benefit of using models with IVs when the IVs are weak albeit the estimators have higher variances. The S–shape IV estimator has a smaller RMSE median and variance than the LL IV estimator, which indicates that shape constraints are helpful to improve performance and robustness of the estimator. Even if we only impose the shape constraints on the last iteration of the Landweber–Fridman iteration, we observe a significant improvement relative to the unconstrained estimator. As the sample size increases, the difference between S–shape IV and LL IV estimator becomes small due to the convergence of LL IV estimator. In contrast, for the incorrectly specified estimators, the performance of the estimators do not significantly improve even if we increase the sample size due to model misspecification.

4.3.3 Growing industry

We assume that the true production function has increasing returns to scale followed by decreasing returns to scale. We consider a similar DGP as the one presented in Olesen and Ruggiero (2014).

$$g_0\left(X_L, X_K\right) = F\left(G\left(X_L, X_K\right)\right) \tag{4.15}$$

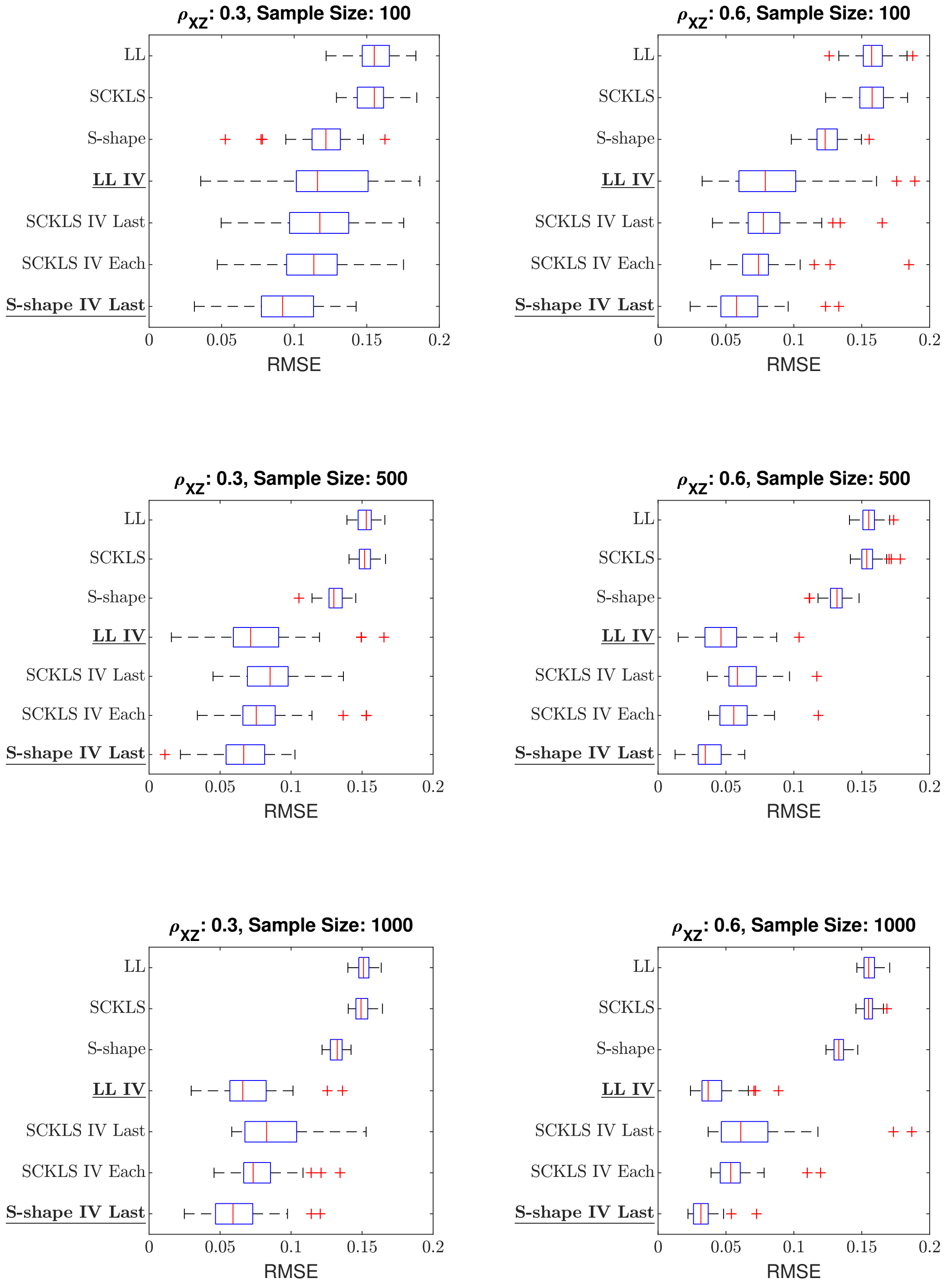


Figure 4.2. The results with an emerging industry DGP

where

$$\begin{aligned}
&G(X_L, X_K) = \left(\beta X_L^{\frac{\sigma-1}{\sigma}} + (1-\beta) X_K^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}} \\
&F(w) = \frac{15}{1+\exp(-4\log 2w)} \\
&\beta = 0.45, \sigma = 1.51
\end{aligned}$$

$G\left(\cdot\right)$ is called a core function defining the input aggregation between labor and capital. The definition of $G\left(\cdot\right)$ affects the shape of input isoquants. $F\left(\cdot\right)$ is called a scale function defining the shape of production function along a ray from the origin.

Figure 4.3 shows the box plot of RMSE values of each estimator for a growing industry. For this industry, only the LL with IV and the S-shape with IV estimators have correctly specified underlying models.

The results are similar to those for the emerging industry case. S–shape IV estimator reduces the median RMSE and variance particularly when the data sample size is small. In contrast to the emerging industry results, the SCKLS IV estimator performance improves. This result is due to the true production function in the growing industry case having a region which is concave consistent with the assumptions of SCKLS. Specifically, the production function beyond the inflection point is correctly specified for the SCKLS estimator, thus the SCKLS estimator performance is improved.

4.3.4 Matured industry

We assume that the true production function is Cobb–Douglas with decreasing returns to scale.

$$g_0\left(X_L, X_K\right) = X_L^{0.4} X_K^{0.4}. \tag{4.16}$$

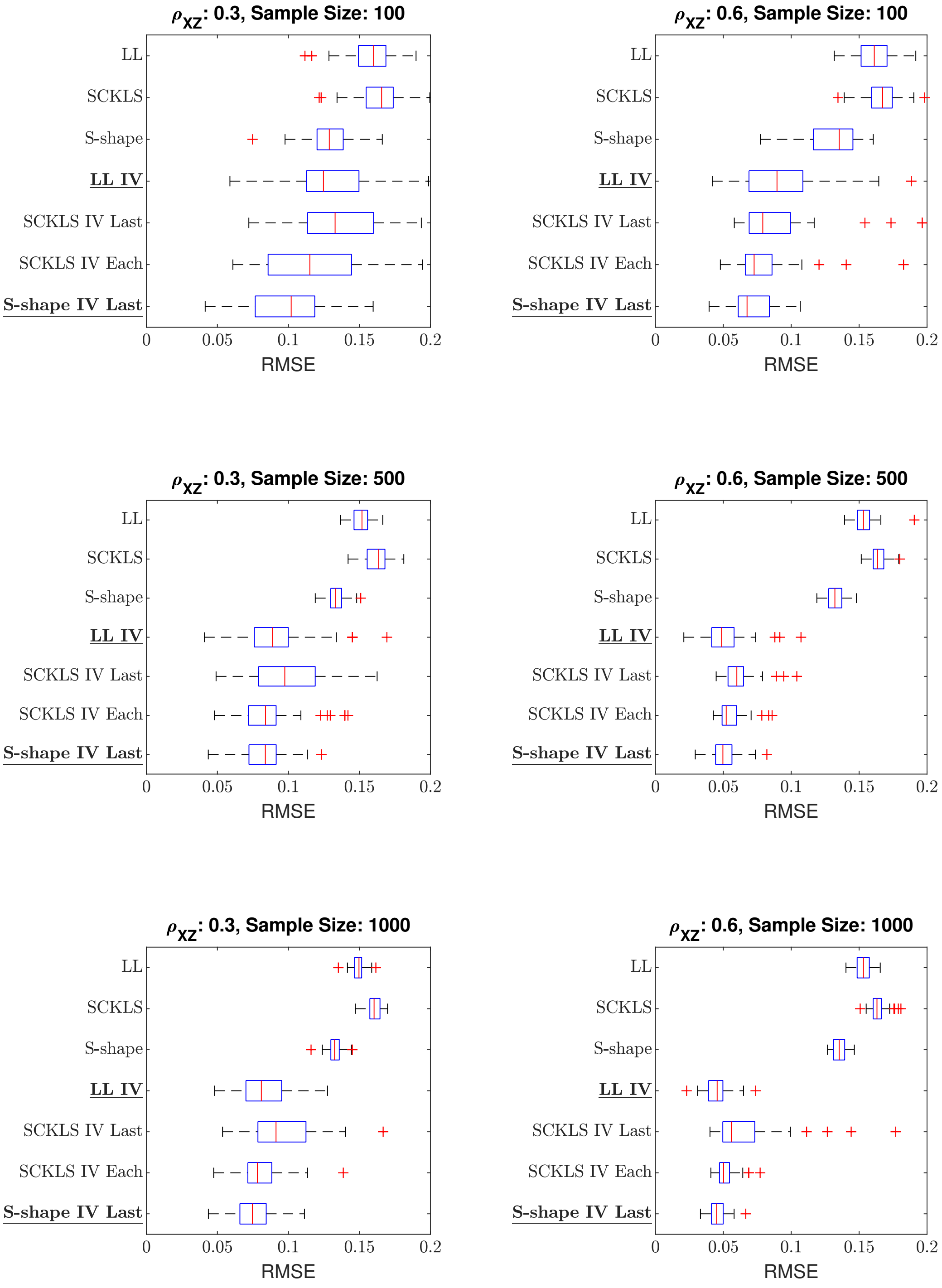


Figure 4.3. The results with a growing industry DGP

Again, we recognize the selection of coefficients of Cobb–Douglas function seems small. We purpose is to demonstrate the robustness of our estimators even in extreme scenarios.

Figure 4.4 shows the estimators performance over 50 simulations using box–plots and the RMSE metric for a matured industry. Here, all IV estimators are correctly specified. Note that the S–shape estimator can be globally concave by setting the inflection point to the origin.

The result shows that shape constrained IV estimators have very competitive performance. All of shape constrained estimators have smaller median RMSE and variance than the unconstrained IV estimator. The benefits are particularly pronounced when the sample size is small. The benefits of imposing shape constraints at each iteration of the SCKLS IV method are slight. Therefore, we recommend imposing the shape constraints only at the last iteration to decrease the computational time.

4.4 Application

We estimate a set of production functions using firm–level data from Japan's *Census of Manufactures* provided by METI. As Foster et al. (2008) suggested, we analyze manufacturing industries which produce homogeneous physical output with small quality variation and small technology changes over time. We select the following seven industries: raw cane sugar (sugar), bread, coffee, plywood, corrugated cardboard box (cardboard), ready-mixed concrete and concrete products. We compare conventional parametric IV estimation with the proposed nonparametric shape constrained IV estimator, and obtain economic insights regarding productivity growth and most productivity scale size of each industry.

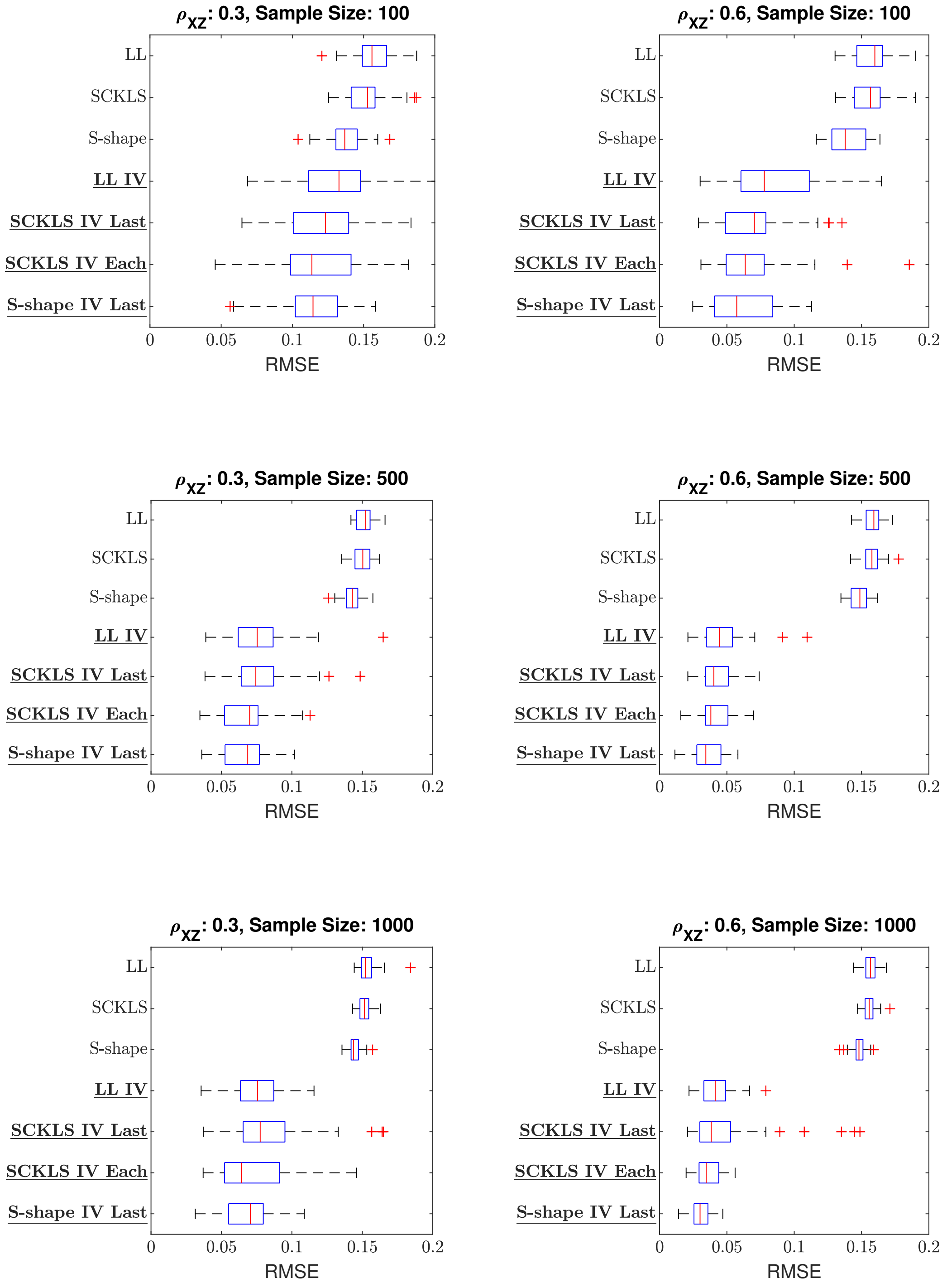


Figure 4.4. The results with a matured industry DGP

4.4.1 Census of Manufactures, Japan

The annual *Census of Manufactures* covers all establishments with four or more employees and is conducted by METI under the Japanese Statistics Act. We use establishment-level data with 30 or more employees since the establishment with less than 30 employees do not report capital stock values. We use the same definition of the variables for production functions as Ichimura et al. (2011):

- L_{jt} = (sum of total regular employees[3] at the end of each month)
- K_{jt} = (starting amount of tangible assets[4])
- y_{jt} = (total amount shipped) + (ending inventory of finished and work-in-progress products) - (starting inventory of finished and work-in-progress products) - (cost for intermediate inputs[5])

where L_{jt},K_{jt} and y_{jt} indicate the labor, capital and value added of firm $j = 1, \ldots, n$ at time $t = 1, \ldots, T$. Then, the production function is modeled as $y_{jt} = g_0(L_{jt}, K_{jt})$ for each industry.

We propose to use the number of full–time labors as the instrumental variables. While managers cannot change full–time labor easily due to the restrictions from the labor union, they can flexibly adjust labor input by hiring or firing part–time employees. There are other potential IVs such as input price. However, input prices have very little variation across firms in this set of homogeneous product industries. Furthermore, input prices are highly correlation with capital which is expected to be a exogenous variable. This correlation is due to the fact that highly capitalized firms are well–organized and paying higher salary to employees.

[3] Regular employees include full-time, part-time, and dispatched workers who work 18 days or more per month.

[4] Tangible assets include machines, buildings, and vehicles.

[5] Intermediate inputs include raw materials, fuel and electricity.

The use of control functions is another promising approach for the endogeneity problem by predicting the productivity using the amount of investment (Olley and Pakes (1996)) or intermediate input (Levinsohn and Petrin (2003)). However, both of these variables potentially contain much larger measurement errors than the headcount of full–time labors. Based on these observations, we propose to implement our IV approach using the number of full–time labors as an instrumental variable. Figure 4.5 shows the transition of full–time and part–time labor over time for each industry. We observe that firms tend to adjust the part–time labor, which is likely to be a cause the endogeneity issue. The fluctuations of other variables over the observation period are described in Appendix C.1.

We use industry–level deflators obtained from the Japan Industrial Productivity Database (JIP)[6] to convert nominal values of capital and value added into real values in 2000. Then we convert establishment-level data into firm-level data by summing up the establishment-level data which belong to the same firm. We use firm-level data because adjustments to the endogenous variable (labor) are typically made at the firm-level by based on the firms productivity level.

Table 4.1 shows the summary statistics of the observed data for each industry. Note that we normalized each variable by dividing the variable by its standard deviation.[7] There are relatively few firms in sugar and coffee industries while the median of value added is larger and capital intensity greater than other industries. This indicates that these two industries contain a few dominant firms which produce most of the demand. The correlation between the headcount of total labors and full-time labors is fairly high but not perfectly correlated, which indicates the firms may change the labor input by adjusting the number of part–time employees. Ready-mix concrete

[6] The JIP database is publicly available at *Research Institute of Economy, Trade and Industry* (REITI) (`https://www.rieti.go.jp/en/database/jip.html`)

[7] This is particular important for the iterative S–shape estimator because it is sensitive to the scale difference between variables.

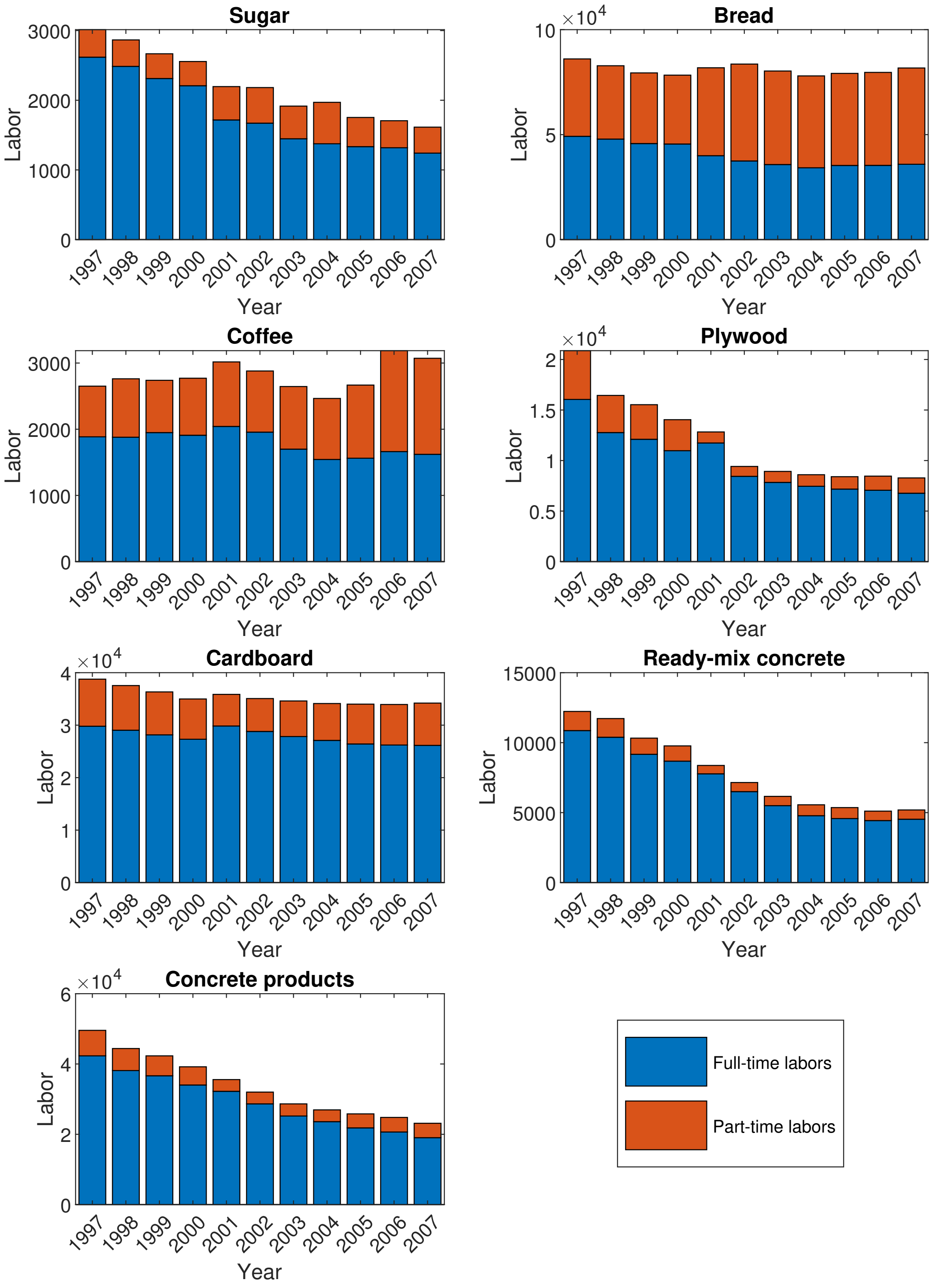


Figure 4.5. Transition of labor input

has a particularly small correlation value between capital and instruments (full–time labor). The low correlation is likely caused by large fluctuations in the industry that are covered by adjusting part–time labor.

Table 4.1. Summary statistics of homogeneous product industries

Industry	Sample size	Median of Input ratio (K/L)	Median of Value Added (y)	ρ_{LZ}	ρ_{KZ}
Sugar	249	1245.9	100036.0	0.984	0.832
Bread	3681	197.8	48773.5	0.992	0.984
Coffee	378	789.8	61991.7	0.923	0.783
Plywood	1495	364.3	58362.1	0.979	0.625
Cardboard	5844	519.8	59305.6	0.983	0.936
Ready-mix concrete	1787	404.5	37602.5	0.962	0.274
Concrete products	5916	452.3	51471.6	0.979	0.645

Before estimating the production functions, we conduct hypothesis tests for correct parametric functional specification: OLS and 2SLS with Cobb–Douglas.[8] We use the parametric specification test described in Henderson and Parmeter (2015).

Table 4.2 shows the p-value for each industry obtained using a bootstrap sample $B = 1000$. The Cobb–Douglas production function cannot be rejected for coffee industry. However, the coffee industry has a small sample of firms and the standard errors of coefficients are large. The Cobb–Douglas 2SLS estimator cannot be rejected for the cardboard and concrete industries. However, for other industries, these parametric specifications are likely to be violated as indicated by the low p-values for the tests. Therefore, more flexible estimators are preferred to avoid functional form

[8] We obtain $\hat{L}$ by estimating the first stage with OLS: $L = \gamma_0 + \gamma_1 Z + v$. Then we obtain production function by estimating second stage with Cobb-Douglas: $\ln(y) = \beta_0 + \beta_1 \ln(\hat{L}) + \beta_2 \ln(K) + \epsilon$.

misspecification.

Table 4.2. Parametric specification test results

Industry	p-value	
	OLS	2SLS
Sugar	0.041**	0.045**
Bread	0.026**	0.031**
Coffee	0.127	0.042**
Plywood	0.015**	0.073*
Cardboard	0.000***	0.207
Ready-mix concrete	0.031**	0.004***
Concrete product	0.060*	0.205

4.4.2 Estimation of production functions and interpretations

4.4.2.1 The setup

Based on our testing results, our primary model for each industry will be as follows: Coffee – OLS, Cardboard and Concrete products – 2SLS, and for all others – the S-shape estimator with IV. We will estimate all three models (OLS, 2SLS, and the S-shape estimator with IV) and the S-shape estimator without IV for all industries help understand the differences for each estimator.

For the iterative S–shape algorithm, we define five rays and five isoquant y–levels by taking 10th, 30th, 50th, 70th and 90th percentile of capital/labor ratio $\{K_{jt}/L_{jt}\}_{j=1,t=1}^{n,T}$ to specify the rays and the same percentiles of value added $\{y_{jt}\}_{j=1,t=1}^{n,T}$ to specify the isoquants. We perform 50 iterates of the S–shape algorithm setting the bandwidth between rays to 0.25 and increasing the bandwidth by 0.25 in each iteration. For Landweber–Fridman regularization, we use relaxation parameter $c = 0.5$ and stopping criteria defined in the equation 4.10 with $\delta = 0.01$.

4.4.2.2 Productivity analysis

We estimate the productivity–level using the estimated production functions. Productivity is measured as the ratio of observed value added y_{jt} to the estimated production function evaluated at the corresponding input factors, $\{L_{jt}, K_{jt}\}$. Intuitively, firms have higher productivity when they can produce larger value added with a given amount of input factors. Productivity can be seen as the firms' deviation of output which cannot be explained by the input factors: managerial practices, quality of input factors, R&D, learning by doing, product innovation, firms' structure decision, or other external drivers, see Syverson (2011) for more details.

To analyze the industry–level productivity transition over time, we aggregate the productivity by computing weighting average with the firms' share of industry output as described in Levinsohn and Petrin (1999). Specifically, we compute the aggregated productivity $\hat{\omega}_t$ for each industry,

$$\hat{\omega}_t = \sum_{j=1}^{n} s_{jt}\hat{\omega}_{jt},$$

where $\hat{\omega}_{jt} = \frac{y_{jt}}{\hat{g}_0(L_{jt},K_{jt})}$ is the estimated firm–level productivity and $s_{jt} = \frac{O_{jt}}{\sum_i^n O_{it}}$ is the firms' share of industry output.[9]

Figure 4.7 shows the percentage growth of the aggregated productivity of each homogeneous product industry estimated by Cobb–Douglas 2SLS, S–shape with IV and without IV.[10] We first see that the sugar and coffee industries are much more volatile than other industries. This due to the small sample size of firms in these two industries and the productivity estimates likely contain significant noise even after aggregation.

[9] Amount of output O_{jt} is the sum of amount shipped and amount of inventory increased during a year.

[10] To avoid having too many information in the figure, we move the estimation results of Cobb–Douglas OLS in Appendix C.1.

The bread industry has significant different productivity estimates between the parametric and the nonparametric models. This can be seen as evidence of parametric misspecification. While Cobb–Douglas 2SLS indicates that aggregated productivity is declining over time, the S–shape estimator shows that productivity is increasing over time. We point to this as evidence that misspecification of the production function's functional form may result in completely opposite conclusions. We also find that S-shape without IV model underestimates the productivity level compared with S-shape with IV for the bread industry. Figure 4.5 shows the labor transition over time for bread industry. We observe that the total labor level is stable while the ratio of part–time labors to full-time labors is increasing over time. Since firms' labor adjustments are not considered in the S–shape without IV model, it leads to the bias in the aggregated productivity estimates.

The productivity level of cardboard and plywood industries are slightly increasing over time. Although in 2007, the plywood industry experiences a significant drop in productivity due to the economic crisis and declining investment in housing. Further, there are significant difference between Cobb–Douglas 2SLS and other estimators in plywood industry. Based on the hypothesis testing results in Table 4.2, we conclude that the Cobb–Douglas 2SLS misspecifies the functional form and overestimates the productivity level.

Ready–mix concrete and concrete products industries have a similar trend of the growth of productivity. We observe the the transition of concrete products industry happens one year after the transition of ready–mix concrete industry. Ready–mix concrete is more sensitive to the demand since it is not possible to store ready–mix concrete while concrete products can be stored as inventory after manufacturing. Furthermore, ready–mix concrete is sensitive to market cycles because large–scale construction, such as bridges, dams and roads, is often postponed in times of economic downturns.

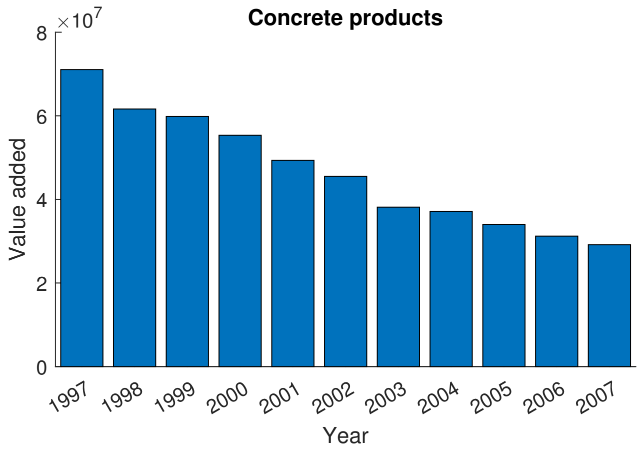


Figure 4.6. Transition of value added in the concrete products industry

We analyze the productivity changes in the concrete products industry by analyzing productivity changes based on the firm's scale size. Figure 4.8 groups firms based on their output levels in 1997 and shows the transition in productivity from 1997-2007. As shown in Figure 4.6 which shows the transition of value added, concrete products industry itself is declining since 1997. However, Figure 4.8 shows that larger firms tend to increase the productivity level while smaller firms are declining. This indicates that sector output tended to be reallocated from small firms to larger firms over the 1997-2007 period. This is explained by the fact that concrete products industry are heavily dependent on capital and most firms are operating well below most productive scale size, Figure 4.9. Thus, it is difficult for smaller firms to improve productivity endogenously over the observed time period.

4.4.2.3 Most productive scale size

We also analyze the optimal scale for each industry by computing the most productive scale size (MPSS) which is the scale size maximizing the average product. Note that MPSS is not well–defined with Cobb–Douglas 2SLS model since MPSS becomes just either zero or infinite

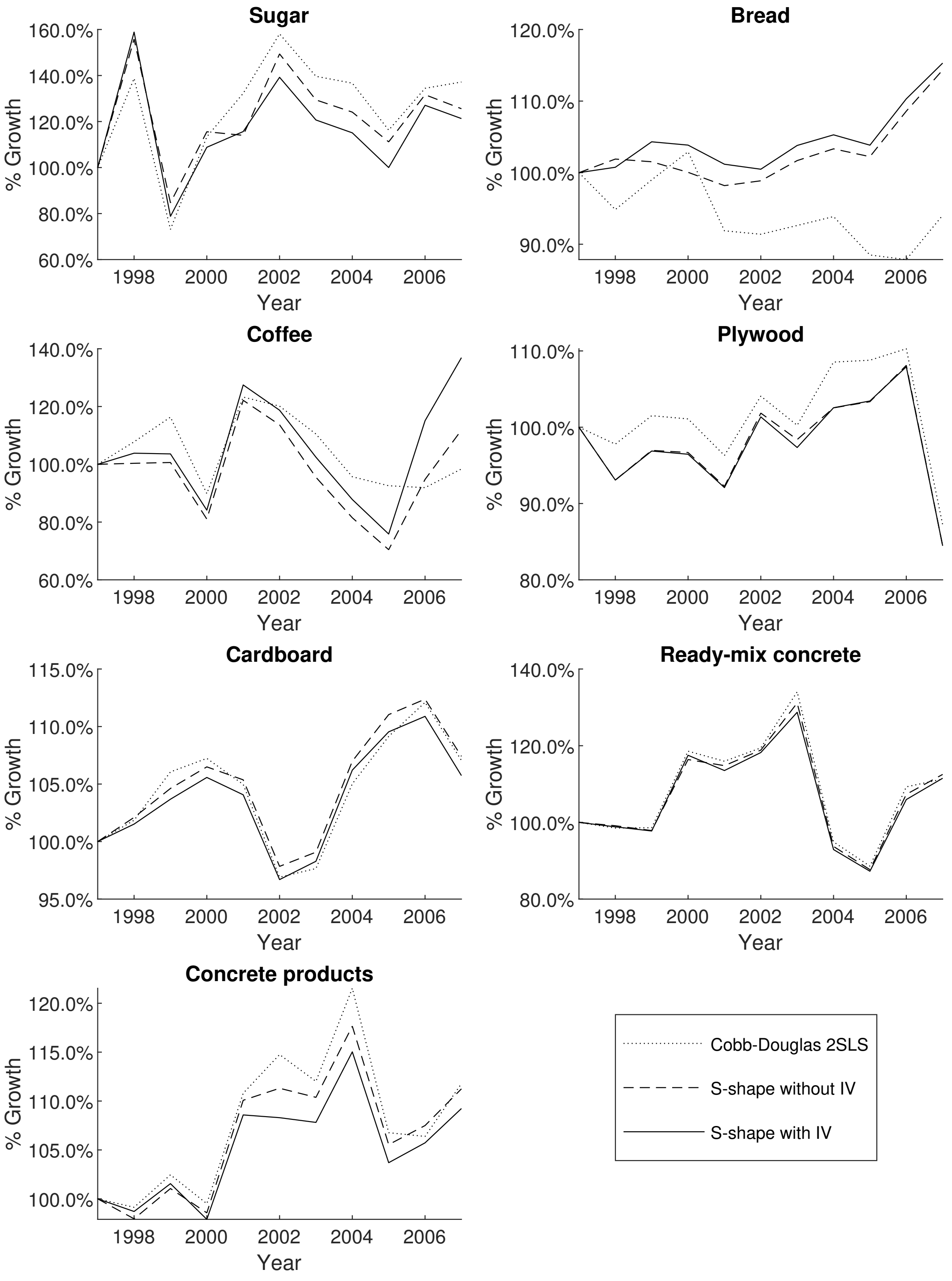


Figure 4.7. Percentage growth of the aggregated productivity from 1997

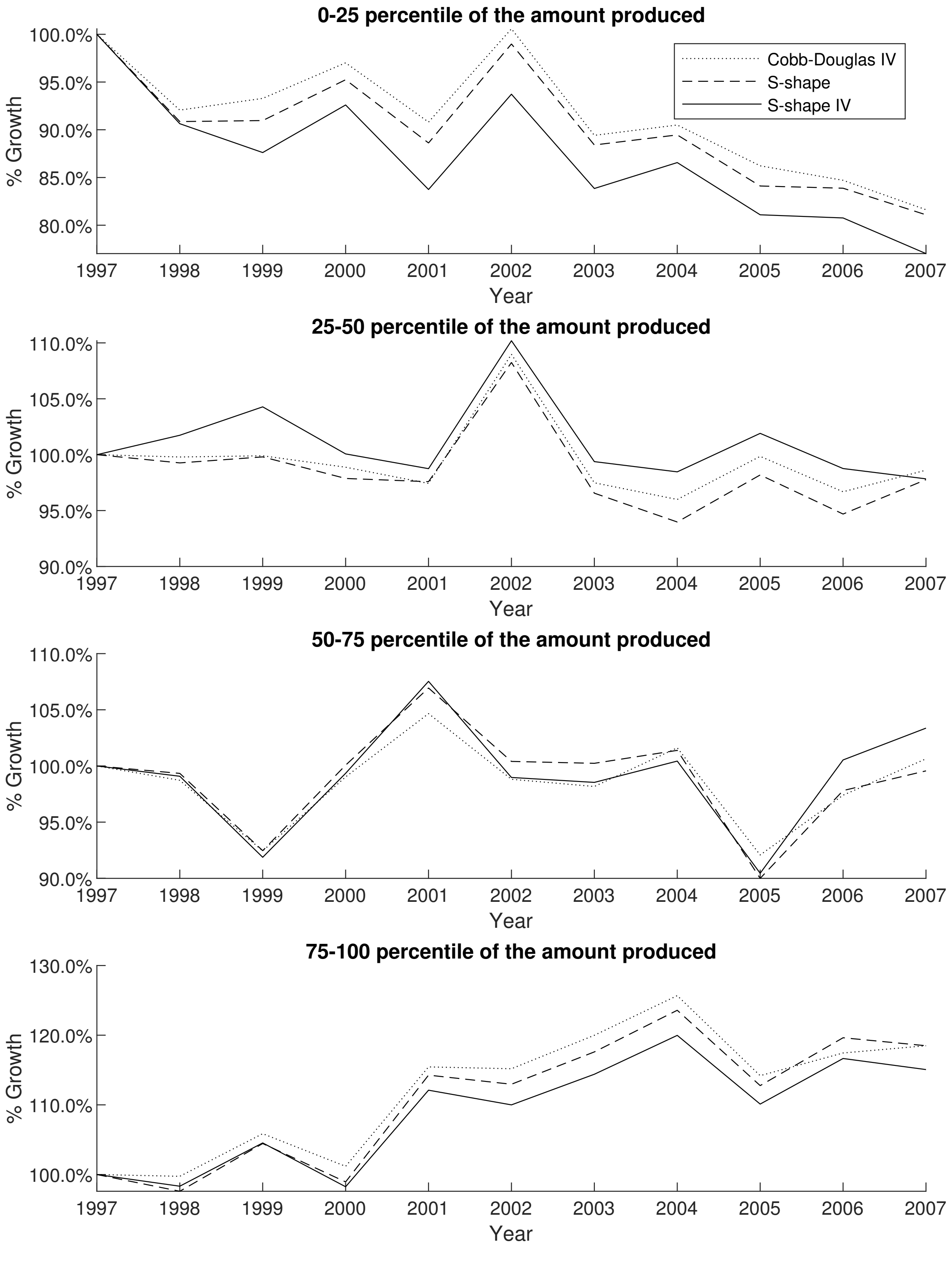


Figure 4.8. Percentage growth of quartile mean of productivity of concrete products industry

depending on returns to scale. Figure 4.9 shows the estimated MPSS of two industries: bread and concrete products with the S–shape IV estimator. Five dotted lines indicate 10th, 30th, 50th, 70th and 90th percentile of capital/labor ratio respectively. Appendix C.1 describes the comprehensive analysis of MPSS for all homogeneous products industries.

For the bread industry, we observe that there is the dominant firm which is operated with the scale much larger than the MPSS. Since this firm controls a large part of the market, they can increase their profitability even if they reduce the productivity by expanding beyond the MPSS. This could be optimal behavior if nonlinear pricing is practiced in the bread industry. However, note lower capital/labor ratio rays have much smaller MPSS estimates, and most labor intensive firms are operating below MPSS. This indicates that labor intensive firms are not able to improve the productivity just by increasing the scale size while maintaining the current input ratio. Instead, they need to invest on capital and automate manufacturing process to increase their scale size.

The concrete products industry does not have a dominant firm although there are a few relatively larger firms in the market. We observe that generally lower capital/labor ratio lines have smaller MPSS estimates. This can be explained by the economic theory that firms become more capital intensive as they grow larger by automating manufacturing processes. Also note that the highest capital/labor ratio line has a slightly smaller MPSS value than the neighbor. This indicates some firms have over–invested in capital and could reduce their capital/labor ratio and improve productivity.

4.5 Conclusion

This chapter proposed a shape constrained IV estimator to address the endogeneity issue while maintaining the flexible nature of nonparametric shape constrained estimators. We propose to

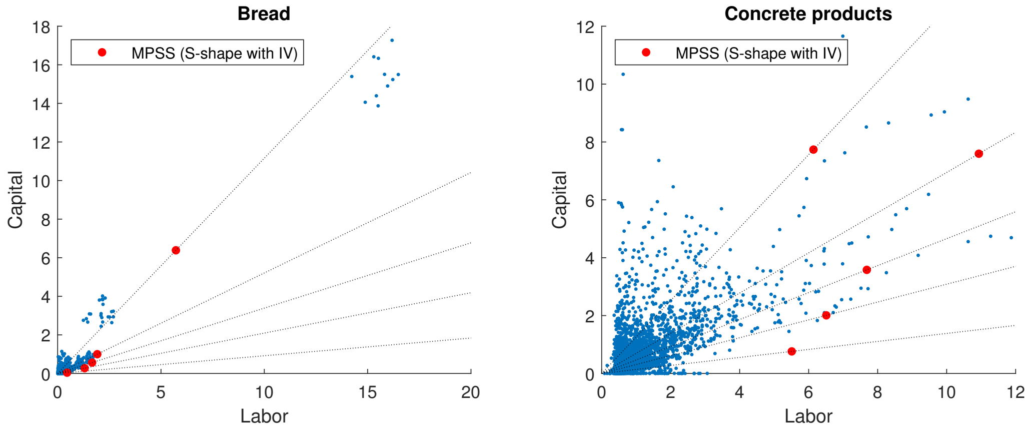


Figure 4.9. Estimated MPSS for bread and concrete products industries

apply shape constraints to estimators used in a Landweber–Fridman iteration regularization algorithm. This algorithm is computation feasible for different types of shape constraints such as global concavity and S–shape.

We validate the finite sample performance of the proposed estimator through simulations under different assumptions on the market conditions. Correctly specified shape constrained model have better performance than unconstrained estimator particularly when sample sizes are small. Furthermore, even with a weak IV, the proposed estimator has significantly better performance than non–IV estimators.

Finally, we apply the proposed estimator to the Japanese census of manufactures to analyze the productivity of homogeneous product industries. We compute the aggregated productivity with three different models, and validate the significant deviations of productivity estimates across model specifications. Specifically, Cobb–Douglas 2SLS may be biased due to the misspecification of the functional form and the S–shape without IV estimator is likely to be biased because of the endogeneity. We also compute the MPSS for each industry to analyze the industry market condition and potential expansion strategy, which cannot be obtained by restrictive parametric models such as 2SLS Cobb–Douglas.

Potential extension of our estimator is to impose S–shape constraints on each Landweber–Fridman iteration to improve the performance at the cost of a more computational intensive algorithm. Also, we plan to clarify the theory of the proposed estimator such as consistency and rate of convergence.

5. CONCLUSIONS

We have developed both models and estimators for production functions with local weighting of data and shape constraints. We developed the SCKLS estimator that imposes simple shape constraints explaining a production process with decreasing returns to scale: global concavity and monotonicity. Then we extended the model to the more general economic restrictions: S-shape expansion and endogeneity.

For the SCKLS estimator, we validated the robustness of the estimator by showing the theoretical properties such as consistency and rate of convergence. We also proposed a test for shape constraints to avoid a misspecification of the shape. Furthermore, we showed the improved finite sample performance via Monte Carlo simulations. Specifically, the SCKLS estimator has better out–of–sample performance than other existing nonparametric estimators even when data samples are small making the SCKLS estimator useful for the analysis of survey data. The SCKLS estimator can also improve the computation efficiency because the size of an optimization problem of the SCKLS estimator is fully controllable by the density of evaluation points, and the simulated performance is stable even with a rough grid of evaluation points.

We extended the SCKLS estimator to describe more general economic axioms allowing smaller firms to benefit from increasing returns to scale through specialization and learning, which is referred to as an S–shape production function. We proposed the iterative algorithm to estimate a production function satisfying the S–shape restriction, convex input sets and allowing for non–homotheticity in input isoquants. In addition to showing the performance improvement through Monte Carlo simulations, we applied the proposed estimator to data from the Japanese census of manufactures. We investigated the supply–side of the cardboard industry and reported most productive scale size and productivity estimates.

Finally, we integrated the use of instrumental variables into our models to address the potential endogeneity of labor. We proposed to apply shape constraints to the Landweber–Fridman regularization process which is computationally less expensive than other regularization methods. We showed the finite sample performance improvement through simulations. Furthermore, we applied the proposed estimator to the Japanese census of manufactures to analyze the productivity of homogeneous products industries. We address the endogeneity of labor by including full–time labor headcount as an instrument for total labor because managers typically cannot adjust full–time labor easily in Japan for unexpected output shocks. We computed most productive scale size and transition of aggregated productivity to reveal the supply–side of homogeneous products industries in Japan.

In summary, this dissertation established the theory and application of local weighted shape constrained estimator by imposing general economic axioms on a production function estimation. Potential future research could focus on the estimation of a production frontier function for efficiency analysis. If systematic inefficiency is present in the data, deconvoluting the residuals, following the stochastic frontier literature, would allow the estimation of a production frontier. Furthermore, we could also extend our models by including firms entry/exit behaviors to describe the industry dynamics over time.

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.

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.

Beattie, B. R., C. R. Taylor, and M. J. Watts (1985). *The economics of production*. Wiley New York.

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*, 513âĂŞ534.

Chambers, R. G., Y. Chung, and R. Färe (1998). Profit, directional distance functions, and nerlovian efficiency. *Journal of Optimization Theory and Applications 98*(2), 351–364.

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, 09). 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.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

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.

Florens, J., J. Racine, and S. Centorrino (2018). Nonparametric instrumental variable derivative estimation. *Journal of Nonparametric Statistics* (just-accepted).

Førsund, F. R. and L. Hjalmarsson (2004). Are all scales optimal in dea? theory and empirical evidence. *Journal of Productivity Analysis 21*(1), 25–48.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *The American economic review 98*(1), 394–425.

Foster, L., J. Haltiwanger, and C. Syverson (2016). The slow growth of new plants: Learning about demand? *Economica 83*(329), 91–129.

Fried, H. O., C. K. Lovell, and S. S. Schmidt (2008). *The measurement of productive efficiency* *and productivity growth*. Oxford University Press.

Frisch, R. (1964). *Theory of production*. Springer.

Ghosal, P. and B. Sen (2016). On univariate convex regression. arXiv preprint arXiv:1608.04167.

Ginsberg, W. (1974). The multiplant firm with increasing returns to scale. *Journal of Economic Theory 9*(3), 283–292.

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 J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics 33*(6), 2904–2929.

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

Haltiwanger, J., R. S. Jarmin, and J. Miranda (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics 95*(2), 347–361.

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.

Horowitz, J. L. (2014). Ill-posed inverse problems in economics. *Annu. Rev. Econ. 6*(1), 21–51.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics 124*(4), 1403–1448.

Hwangbo, H., A. L. Johnson, and Y. Ding (2015). Power curve estimation: Functional estimation imposing the regular ultra passum law. SSRN working paper available at: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2621033`.

Ichimura, H., Y. Konishi, and Y. Nishiyama (2011). An econometric analysis of firm specific productivities: Evidence from japanese plant level data. Discussion papers, Research Institute of Economy, Trade and Industry (RIETI).

Iguchi, M. (2015). Growth strategy of cardboard industry: demand forecast and growth directions. Master's thesis, Waseda University (in Japanese).

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., A. Johnson, and A. Saastamoinen (2015). Stochastic nonparametric approach to efficiency analysis: A unified framework. In *Data Envelopment Analysis*, pp. 191–244. Springer.

Kuosmanen, T. and A. L. Johnson (2017). Modeling joint production of multiple outputs in stoned: Directional distance function approach. *European Journal of Operational Research 262*(2), 792–801.

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 (1999, January). When industries become more productive, do firms? Working Paper 6893, National Bureau of Economic Research.

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.

Marschak, J. and W. H. Andrews (1944). Random simultaneous equations and the theory of production. *Econometrica, Journal of the Econometric Society 12*(3/4), 143–205.

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.

Mundlak, Y. (1961). Empirical production function free of management bias. *Journal of Farm Economics 43*(1), 44–56.

Mundlak, Y. and I. Hoch (1965). Consequences of alternative specifications in estimation of cobb-douglas production functions. *Econometrica: Journal of the Econometric Society 33*(4), 814–828.

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

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603.

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.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*(6), 1263–1297.

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 (Statistical Methodology) 2*(79), 423–448.

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

Syverson, C. (2004). Product substitutability and productivity dispersion. *The Review of Economics and Statistics 86*(2), 534–550.

Syverson, C. (2011). What determines productivity? *Journal of Economic literature 49*(2), 326–365.

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, 12). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*(4), 1261–1295.

Yagi, D., Y. Chen, A. L. Johnson, and T. Kuosmanen (2018). Shape constrained kernel-weighted least squares: Estimating production functions for chilean manufacturing industries. *Journal of Business & Economic Statistics* (just-accepted).

Yagi, D., Y. Chen, A. L. Johnson, and H. Morita (2018). An axiomatic nonparametric production function: Modeling production in japan's cardboard industry. Working paper.

Zellner, A., J. Kmenta, and J. Dreze (1966). Specification and estimation of cobb-douglas production function models. *Econometrica: Journal of the Econometric Society 34*(4), 784–795.

APPENDIX A

APPENDIX OF CHAPTER 2

A.1 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.1.1 More on practical considerations and extensions to SCKLS

A.1.1.1 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.1.1.2 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).

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

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

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

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

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

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

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

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 $p \geq 3$ to improve its convergence performance (at least theoretical). But that would require much stronger assumption on the smoothness of g_0, and would lead to additional computational burden[1]. Our experience suggests that 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.1.1.3 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[2]. 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

[1] While the optimization problem is still polynomial time solvable, the number of decision variables would increase and the constraint matrix would become significantly more dense, lending to computational challenges.

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

leads to a different weighting scheme in the objective function, as illustrated in the following.

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

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).

A.1.1.4 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 $\{\boldsymbol{X}_j\}_{j=1}^n$.

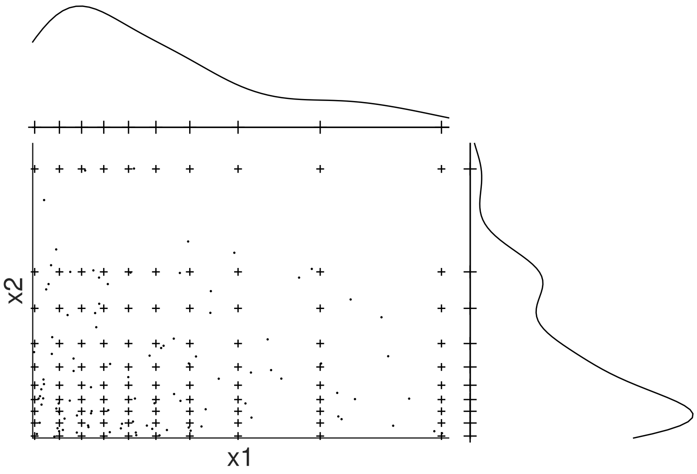


Figure A.1. Example of non-uniform grid with kernel density estimation.

A.1.2 Some related work

A.1.2.1 Convex Nonparametric Least Squares (CNLS)

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

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

by solving the quadratic programming (QP) problem

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

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 $\mathcal{G}^n = \{g : \mathbb{R}^d \to \mathbb{R} \mid g(\boldsymbol{x}) = \min_{j\in\{1\ldots,n\}}(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}),$ with $\boldsymbol{\beta}_j \geq 0$ for $j = 1,\ldots,n\}$. 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.1.2.2 Constrained Weighted Bootstrap (CWB)

A.1.2.2.1 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[4]. The problem is formulated as follows.

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

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

[4] 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.

a two-step process: 1) estimate an unconstrained kernel estimator; 2) find the shape constrained function that is as close as possible (as measured by the Euclidean distance in p-space) to the unconstrained kernel estimator. Based on our experience, CWB tends to suffer from computational difficulties and occasionally poor estimates in small samples. We suggest changing the objective function to minimize the distance from the estimated function to the observed data. This modification seems to improve the estimates empirically as shown in Appendix A.5.

A.1.2.2.2 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.1.2.2.3 Calculating the estimate of the first partial derivative 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 2.5.

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

		Average RMSE on the observation points				
Number of observations		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 $\hat{g}^{(1)}(\boldsymbol{x}|\boldsymbol{p})$.

		Average RMSE on the evaluation points				
Number of observations		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.1.3 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 a weighted function of estimated errors, and therefore avoids overfitting the observed data. However, as shown in A.2.2, 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.1.3.1 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.* $\|\boldsymbol{h}\| \to \boldsymbol{0}$ *(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* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

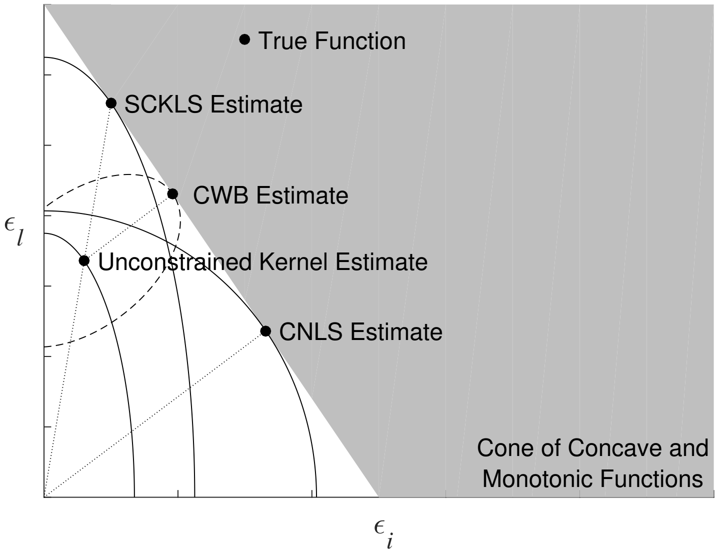


Figure A.2. Comparison of different estimators in the estimated-ϵ-space.

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 2.5 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 2.3 (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 2.1(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.1.3.2 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. $\|\boldsymbol{h}\| \to \mathbf{0}$ (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 $\|\boldsymbol{h}\| \to \mathbf{0}$. Combining with Proposition A.1, CNLS can be viewed as a special case of SCKLS and CWB in y-space.

A.1.3.3 The relationship between CWB in p-space and SCKLS

Again start from the SCKLS estimator, and in view of Assumption 2.1 (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

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

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

Alternatively now consider the objective function of CWB, specifically $D(\boldsymbol{p}) = \sum_{j=1}^{n}(p_u - p_j)^2 = \sum_{j=1}^{n}(1/n - p_j)^2 = L(m(g(p_j)))$. 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.1.3.4 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	$m(m-1)$	$n(n-1)$	$m(m-1)$

A.2 Technical proofs

A.2.1 Summary of the proof strategy

Theorems 2.1– 2.4 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

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

   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 2.1). 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 A.1 and Lemma A.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 2.2). Building upon Theorem 2.1, 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 A.3 and Lemma A.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 2.3). 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 2.4). This can be viewed as a generalization of Theorem 2.2. 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 A.4.

A.2.2 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 $i = 1, \dots, m$, where $(\tilde{a}_1, \tilde{\boldsymbol{b}}_1, \dots, \tilde{a}_m, \tilde{\boldsymbol{b}}_m)$ 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. $\boldsymbol{h} = (h, \dots, h)'$. Since the objective function is quadratic, for any $(a_1, \boldsymbol{b}_1, \dots, a_m, \boldsymbol{b}_m)$, its value equals

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

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

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

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

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

then SCKLS belongs to[5]

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

where G_2 is the set that contains all the concave and increasing functions from $\boldsymbol{S}$ 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 A.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 A.2 (Proposition 7 of Fan and Guerre (2016), Page 509)**.** *Suppose that Assumption 1(i)-1(vi) hold, then as $n \to \infty$,*

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

[5] 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.

A.2.3 Proof of Theorems in Section 2.3

A.2.3.1 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 $i = 1, \dots, m$ is established because our objective function corresponds to is a quadratic programming problem with a positive definite (strictly convex) objective function with a feasible solution. See Bertsekas (1995).

Based on our characterization of SCKLS in Appendix A.2.2, we note that the objective function at the SCKLS estimate is smaller than or equal to that at the truth, and thus

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

Moreover, by the triangular inequality, we have that

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

As such,

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

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

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

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

□

A.2.3.2 Proof of Theorem 2

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

A.2.3.2.1 Some useful lemmas Here we list two useful lemmas on the convergence of convex functions.

Lemma A.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

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

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,

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

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

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

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 A.4 can be viewed as a small extension of Lemma A.3. This is the version that we shall use in the proof of Theorem 2.2.

Lemma A.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 A.3, we see that $\sup_{\boldsymbol{x}\in\boldsymbol{C}} |f_n(\boldsymbol{x}) - f_0(\boldsymbol{x})| \geq \epsilon$ would entail

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

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

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

□

A.2.3.2.2 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 A.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* $\boldsymbol{C} \subset \mathrm{int}(\boldsymbol{S})$ *(where* $\mathrm{int}(\cdot)$ *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

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

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 2.2(i)). Here the last line also follows from Assumption 2.2(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). □

A.2.3.2.3 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 $\boldsymbol{C} \subset \mathrm{int}(\boldsymbol{C}')$ and $\boldsymbol{C}' \subset \mathrm{int}(\boldsymbol{S})$, where $\mathrm{int}(\cdot)$ denotes the interior of a set.

By Lemma A.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 2.1 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 A.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 A.1 and Theorem 2.1,

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

where the first inequality is from Lemma A.1, and the last equality is from Theorem 2.1.

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 $i=1,\ldots,m$. □

A.2.3.3 Proof of Theorem 3

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

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

as $n\to\infty$. It then follows from $h=O(n^{-1/(4+d)})$ and Assumption 2.3 that

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

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 $\boldsymbol{S}$.

□

A.2.4 Proof of Theorems in Section 2.4

A.2.4.1 Proof of Theorem 4

Proof. Using the definition of SCKLS in Appendix A.2.2 and the notation in the proofs of Theorem 1 and Theorem 2, we have that

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

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 A.2, with probability one, for sufficiently large n,

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

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

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 $\mathrm{int}(\boldsymbol{S})$) 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 A.1 and Assumption 2.2(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

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

so (A.14) is no greater than

$$
\begin{aligned}
&2\sqrt{\frac{1}{m}\sum_{i=1}^{m}\big(\tilde{a}_i - g_0^*(\boldsymbol{x}_i), (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))'h\big)\boldsymbol{\Sigma}_i \begin{pmatrix} \tilde{a}_i - g_0^*(\boldsymbol{x}_i) \\ (\tilde{\boldsymbol{b}}_i - \frac{\partial g_0^*}{\partial \boldsymbol{x}}(\boldsymbol{x}_i))h \end{pmatrix}} \\
&\to 2\Big\{\int_{\boldsymbol{S}}(g_0(\boldsymbol{x}) - g_0^*(\boldsymbol{x}))^2 Q(d\boldsymbol{x})\Big\}^{1/2} \leq 2\Big\{\int_{\boldsymbol{S}} g_0^2(\boldsymbol{x})Q(d\boldsymbol{x})\Big\}^{1/2}.
\end{aligned}
$$

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 A.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 A.1 implies that

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

in probability. As such, we can instead work on

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

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 A.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 A.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 $|(\text{II})|$ and $|(\text{III})|$ 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 2.2 via Lemma A.4.

□

A.2.4.2 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 A.2.2, 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 $\mathbf{1}_{\{T_n \le T_{nk}^\circ\}} \le \mathbf{1}_{\{T_n \le T_{nk}+\Delta_n\}}$ for large n. Note that

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

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 $\sup_j |\tilde{\epsilon_j} - \epsilon_j| \le O_p(n^{-2/(4+d)} \log^{1/2} n)$. 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 $\mathbf{1}_{\{T_n^\circ \le T_{nk}^\circ\}} \le \mathbf{1}_{\{T_n \le T_{nk}+\Delta_n\}}$ and thus $p_n \ge p_n^\circ$. As a result, $P(p_n \le \alpha) \le P(p_n^\circ \le \alpha) \le \alpha$, 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

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

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

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

for sufficiently large n. This also leads to $\mathbf{1}_{\{T^{\circ}_n \leq T^{\circ}_{nk}\}} \leq \mathbf{1}_{\{T_n \leq T_{nk} + \Delta_n\}}$. 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 2.2(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 $\boldsymbol{S}' \subset \operatorname{int}(\boldsymbol{S})$ such that $Q(\boldsymbol{S}') > 0$ and

$$\inf_{\boldsymbol{x} \in \boldsymbol{S}'} |g^*_0(\boldsymbol{x}) - g_0(\boldsymbol{x})| > \delta.$$

Note that

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

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 2.4. 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 2.4) 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 2.2(ii), since $g_0 \notin G_2$ and the evaluation points are reasonably well spread across $\boldsymbol{S}$ (i.e. Assumption 2.2(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 $n \to \infty$,

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

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

□

A.2.5 Proof of Propositions in Appendix A.1.3

A.2.5.1 Proof of Proposition A.1

Proof. In view of Assumption 2.1 (v), for any sufficiently small $\boldsymbol{h}$, we have

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

Then, the objective function of (2.3) is equal to $\sum_{j=1}^{n}(y_j - a_j)^2 K(\mathbf{0})$, and thus

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

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 (2.3) can be rewritten as follows:

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

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

A.2.5.2 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 (2.3) 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

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

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 (2.3) can be rewritten as follows with (A.17).

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

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

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

□

A.2.5.3 Proof of Proposition A.3

Proof. In view of Assumption 2.1 (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 $\sum_{j=1}^{n}(y_j - a_j)^2K(\mathbf{0})$, and thus

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

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

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

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

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

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

A.3 Testing for affinity using SCKLS

A.3.1 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

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

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

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

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 2.4.2 (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^L_{nk}. Fitting the data using SCKLS (with the same set of evaluation points and the same bandwidth as before) leads to the resulting estimators $\hat{g}^{\mathrm{V}}_{nk}$ and $\hat{g}^{\Lambda}_{nk}$, where $\hat{g}^{\mathrm{V}}_{nk}$ is the SCKLS estimator using only the convexity constraint, while $\hat{g}^{\Lambda}_{nk}$ is the SCKLS estimator using only the concavity constraint, all based on $\{\boldsymbol{X}_j, y_{jk}\}_{j=1}^n$. So

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

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, $\hat{g}^{\mathrm{V}}_n$ and $\hat{g}^{\Lambda}_n$, should be close to the linear fit g^L_n, 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.

A.3.2 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.

A.4 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 select-

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

ing 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.

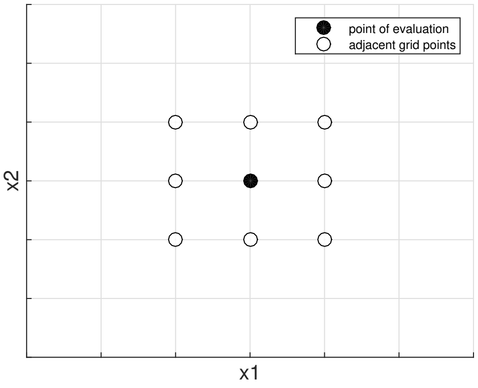


Figure A.3. Definition of adjacent grid in two-dimensional case.

Algorithm 2C Iterative approach for SCKLS computational speedup

```
$t \Leftarrow 0$
$V \Leftarrow \{(i,l) : \boldsymbol{x}_i$ and $\boldsymbol{x}_l$ 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 (2.3) **do**
    $t \Leftarrow t + 1$
    $U \Leftarrow \{(i,l) : \boldsymbol{x_i}$ and $\boldsymbol{x}_l$ do not satisfy constraints in (2.3)$\}$
    $V \Leftarrow V \cup U$
    Solve relaxed SCKLS with $V$ to find solution $\{a_i^{(t)}, \boldsymbol{b}_i^{(t)}\}_{i=1}^m$
**return** $\{a_i^{(t)}, \boldsymbol{b}_i^{(t)}\}_{i=1}^m$
```

A.5 Comprehensive results of existing and additional numerical experiments

We show the comprehensive results of experiments in Section 2.5 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[6].

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[7].

A.5.1 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 2.1 displays the average RMSE of each estimator. The distribution of bandwidths selected by LOOCV

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

[7] 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.1.2.2.

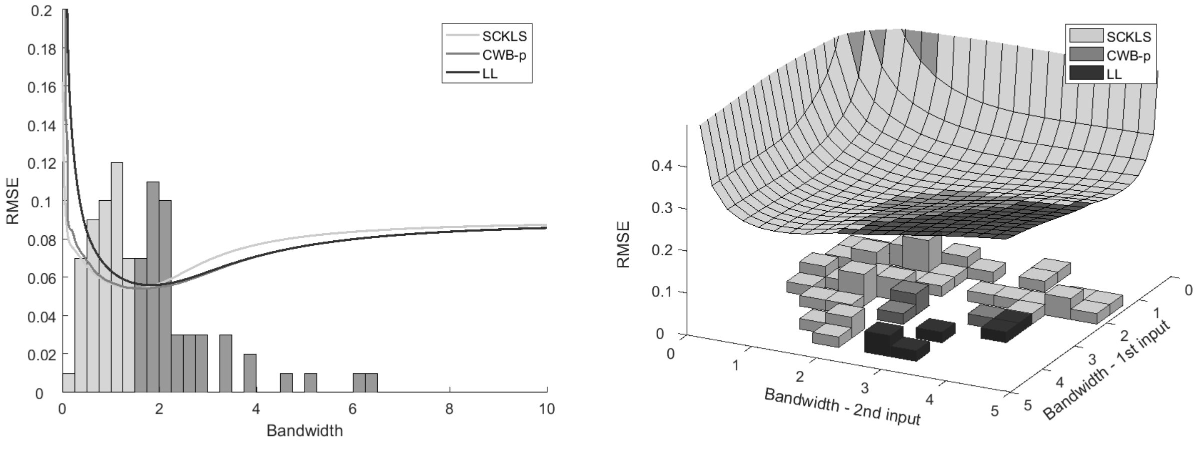
Table A.5. Comprehensive results of RMSE on observation points for Experiment 1

		Average of RMSE on observation points				
	Number of observations	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

are shown in the histogram. The instances 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. Comprehensive results of RMSE on evaluation points for Experiment 1

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

A.5.2 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

		Average of RMSE on observation points				
	Number of observations	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

		Average of RMSE on evaluation points				
	Number of observations	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

A.5.3 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. Comprehensive results of RMSE on observation points for Experiment 2

		Average of RMSE on observation points		
Number of evaluation 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. Comprehensive results of RMSE on evaluation points for Experiment 2

		Average of RMSE on evaluation points		
Number of 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. Comprehensive results of 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%)

A.5.4 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.1.1. 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

		Average of RMSE on observation points				
	Number of observations	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

		Average of RMSE on evaluation points				
	Number of observations	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**

A.5.5 Estimation with a misspecified shape

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

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

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

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

with $\beta = 0.45$ and $\sigma = 1.51$. For $j = 1, \dots, n$, input, $\boldsymbol{X}_j = (X_{j1}, X_{j2})'$, 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

	Average of RMSE on evaluation points				
Number of observations	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

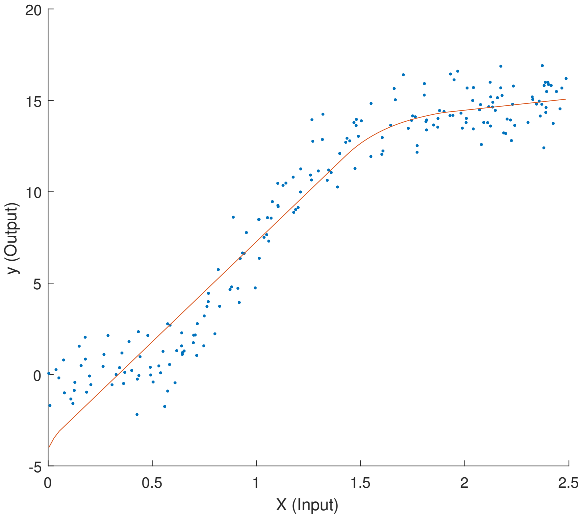


Figure A.5. A typical run of SCKLS when the truth is S-shaped.

A.6 Semiparametric partially linear model

A.6.1 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).

A.6.2 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 2.6 has only two-input, we expect that SCKLS with Z-variables tends not to deteriorate the estimator performance in our application.

A.7 Details on the application to the Chilean manufacturing data

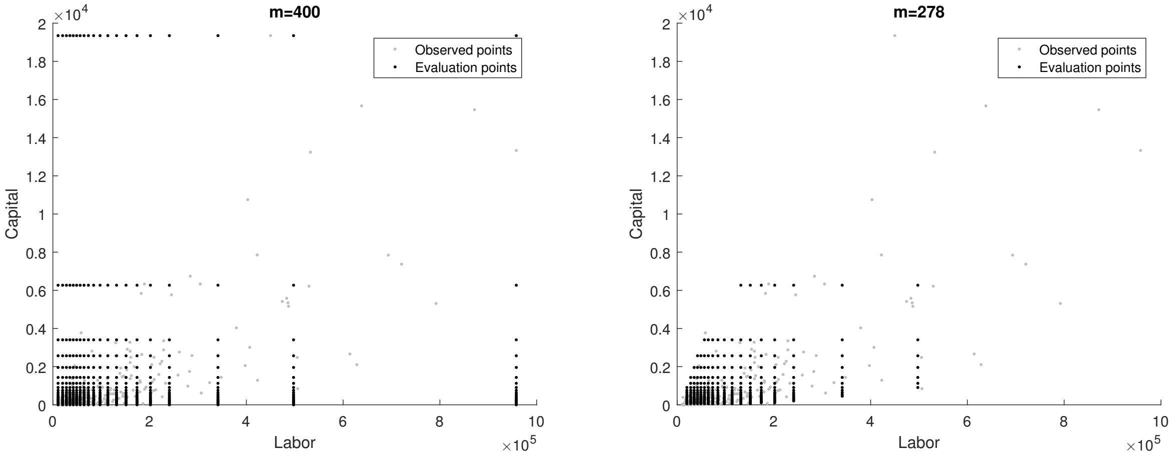
In section 2.6, we applied the SCKLS estimator to the Chilean manufacturing data to estimate a production function for plastic (2520) and wood (2010) industries. Here we provide the detailed specification of the SCKLS estimator applied to the real data. Since the application data is skewed as shown in Table 2.6, 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

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

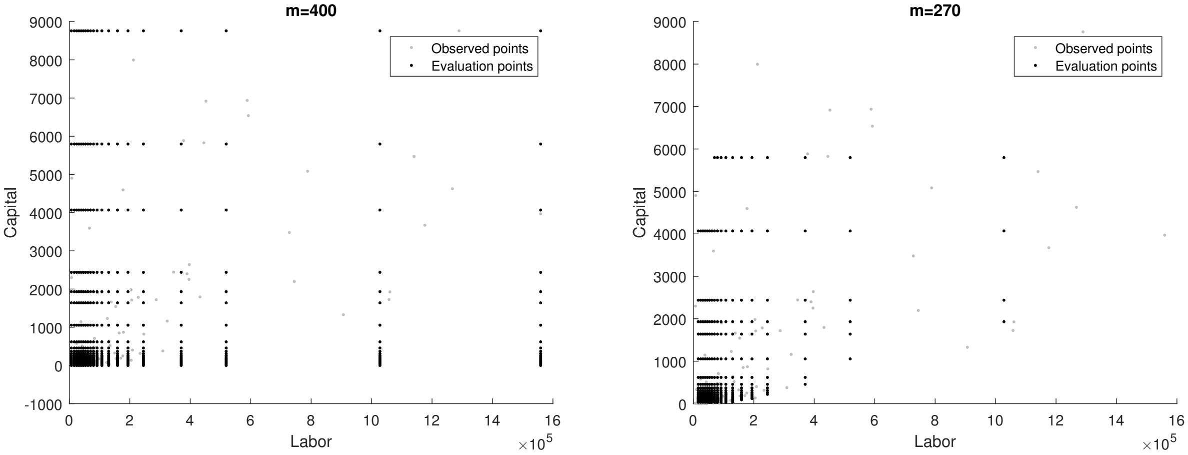


(a) Before deletion

(b) After deletion

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

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.



(a) Before deletion

(b) After deletion

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

APPENDIX B

APPENDIX OF CHAPTER 3

B.1 Technical proofs

B.1.1 Proof of Theorems in Section 3.2

B.1.1.1 Proof of Lemma 3.1

Proof. For simplicity, we focus on the case of $d = 1$. Note that following arguments can be extended for the multiple input case with $d > 1$ by studying the function g_0 along any direction.

Now, the elasticity of scale is defined as

$$\epsilon(x) = g_0'(x)\frac{x}{g_0(x)}.$$

Next we compute the derivative of the elasticity of scale,

$$\epsilon'(x) = \frac{1}{g_0(x)}\left(xg_0''(x) + g_0'(x)\left(1 - \epsilon(x)\right)\right). \tag{B.1}$$

By Definition 3.4, we have following conditions on the elasticity of scale:

$$\epsilon'(x) < 0 \text{ for } \forall x$$

$$\epsilon(x_A) > 1 \text{ and } \epsilon(x_B) < 1 \text{ for some } x_A < x_B.$$

By using these conditions on Equation (B.1) and assumption that g_0 is monotonically increasing, we have,

$$g_0''(x) < 0 \text{ for } \forall x > x_B.$$

Here, by the assumption that there exists a single point of inflection point x^* such that $g_0''(x^*) = 0$, we have

$$g_0''(x) > 0 \text{ for } \forall x < x^*$$

$$g_0''(x) = 0 \text{ for } x = x^*$$

$$g_0''(x) < 0 \text{ for } \forall x > x^*$$

which implies the function $g_0(\cdot)$ is a S-shaped function define in Definition 3.5. □

B.1.1.2 Proof of Lemma 3.2

Proof. First note that the S-shape function is defined for any expansion path and a ray from the origin is a subset of the set of expansion paths. So a single inflection point exist on each 2-D sectional of the production function by definition of an S-shape function. The result to be shown, the set of inflection points lie on the same input isoquant with aggregate input level x_A^*, can be stated mathematically as

$$x_A^* = \underset{x_A \in \alpha \boldsymbol{x}}{\operatorname{argmax}} \left(\frac{dF(x_A)}{dx_A} \,\middle|\, x_A = g(\boldsymbol{x}) \right)$$

for a ray vector $\alpha \boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and the origin define a ray in input space and x_A^* is the inflection point on that ray. By the definition of homothetic we have $f(\boldsymbol{x}) = F(g(\boldsymbol{x}))$. We substitute $x_A = g(\boldsymbol{x})$ the derivative of f with respect to x_A, which is just $\dfrac{dF(x_A)}{dx_A}$. Notice this is independent of the ray from the origin selected. Thus, we have the result. □

B.1.1.3 Proof of Theorem 3.1

Proof. From Lemma 3.2 we know that if the S-shape function definition holds for a ray from the origin then it holds for any ray from the origin and the inflection point will be located on the same isoquant. Now we just need to show for an arbitrary (non-radial) expansion path the that RUP law holds.

By the definition of an expansion path and Lemma 3.2, we see that as we move from input vector $\boldsymbol{X}_{m-1}$ to $\boldsymbol{X}_m$ we move between two input isoquants which are in the same sequential order as they would be for an expansion path along a ray from the origin, thus the passum coefficient is decrease, given us the desired result. □

B.2 Detailed algorithm and estimation procedure

In this section, we described the advanced estimation algorithm and mathematical formulation of each step. The algorithm is constructed by two estimations: (1) input isoquants at some y–levels by Convex Nonparametric Least Squares (CNLS) type estimator, and (2) S-shape functions on some rays from the origin by Shape Constrained Kernel Least Squares (SCKLS). Algorithm 2B presents the details of our advanced algorithm which is composed of three steps: Initialization, Iteration and Updating parameters. The section numbers, where the details of each step are described, are displayed in the right column of the table.

B.2.1 Initialization

The number of isoquants I, the number of rays from the origin R, isoquant y-levels, $y^{(i)}$, and rays from the origin, $\boldsymbol{\theta}^{(r)}$ can be initialized in the same way as what we discussed in section 3.3.3.

In the estimation of S-shape function on rays from the origin, we need to specify the smooth-

Algorithm 2B Details of the advanced estimation algorithm

```
Data: observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$
procedure                                                                  (Section)
Initialization:                                                            (B.2.1)
    $I \leftarrow$ Initialize number of isoquants
    $R \leftarrow$ Initialize number of rays
    $\{y^{(i)}\}_{i=1}^I \leftarrow$ Initialize isoquant $y$-levels
    $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^R \leftarrow$ Initialize rays from origin
    $\boldsymbol{\omega} \leftarrow$ Initialize smoothing parameter between rays
    Project observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ to the isoquant level $y^{(i)}$
    Estimate initial isoquants by the CNLS-based estimation                (B.2.2)
Iteration:
    while Termination condition not reached do
        Project observations onto the ray $\boldsymbol{\theta}^{(r)}$                          (B.2.3.1)
        Update S-shape estimates using the SCKLS-based estimator           (B.2.3.2)
        Update isoquant estimates by the CNLS-based estimator              (B.2.3.3)
        Minimize the gap between S-shape and isoquant estimates            (B.2.3.4)
        Compute Mean Squared Errors against observations                   (B.2.3.5)
Updating parameters:                                                       (B.2.4)
        $I \leftarrow$ Update number of isoquants
        $R \leftarrow$ Update number of rays from the origin
        $\{y^{(i)}\}_{i=1}^I \leftarrow$ Update isoquant $y$-levels
        $\{\boldsymbol{\theta}^{(r)}\}_{r=1}^R \leftarrow$ Update rays from origin
        $\boldsymbol{\omega} \leftarrow$ Update smoothing parameter between rays
    end
return : Estimated function with minimum Mean Squared Errors
```

ing parameter between rays, $\boldsymbol{\omega}$, which determines the weights on each observation based on the angle between the observation and the ray from the origin on which we are currently estimating. Instead of optimizing bandwidth between angels, $\boldsymbol{\omega}$, by grid search in Algorithm 1, we try to find the optimal bandwidth between angles by increasing $\boldsymbol{\omega}$ by some increments, $\Delta\boldsymbol{\omega}$, with updating both isoquants and S-shape estimates in each iteration of Algorithm 2B. Based on our numerical experiments, we recommend to start from a small value and increase $\boldsymbol{\omega}$ by small increment $\Delta\boldsymbol{\omega}$ in each iteration. We will generate a set of estimates and select from the set. Intuitively, the S-shape function, estimated along the ray, only uses observations close to the ray in the first iteration. As an our algorithm progresses, the S-shape estimation step includes data more distance from the ray. For more details of the S-shape estimation and the smoothing parameters, see Appendix B.2.3.2.

B.2.2 Estimate convex isoquants

We are interested in estimating the isoquant function $\mathcal{H}$ in (3.3) at a given level of output. Assume that a set of output levels for isoquant estimation is given by

$$y^{(i)},\ i = 1, \ldots, I \tag{B.2}$$

where I is the number of isoquants to be estimated. Also assume that the input data used to estimate the isoquant at $y^{(i)}$ is given by

$$\boldsymbol{X}^{(i)},\ i = 1, \ldots, I \tag{B.3}$$

where $\boldsymbol{X}^{(i)}$ is subset of observations of input used for the estimation of isoquant at level $y^{(i)}$. $\boldsymbol{X}^{(i)}$ is $n_i \times d$ matrix and n_i denotes the number of observations used for estimation of isoquant i at level $y^{(i)}$. We have already described the procedure for specifying isoquant level $y^{(i)}$ in section 3.3.3 and how to obtain the input data $\boldsymbol{X}^{(i)}$ associated with the isoquant level $y^{(i)}$ in section 3.3.4.1.

It is common that the input data $\boldsymbol{X}^{(i)}$ contains errors in all input dimension $\{\boldsymbol{X}_1^{(i)}, \ldots, \boldsymbol{X}_d^{(i)}\}$ since we just project each observation to the closest isoquant output level. We first propose to use the existing nonparametric estimation method called Convex Nonparametric Least Squares (CNLS) to estimate isoquants. We also propose two modifications to the CNLS estimator which improve the performance of the isoquant estimation.

B.2.2.1 Convex Nonparametric Least Squares (CNLS)

Kuosmanen (2008) extends Hildreth's least squares approach to the multivariate setting with a multivariate $\boldsymbol{x}$, and coins the term Convex Nonparametric Least Squares (CNLS). CNLS builds upon the assumption that the true but unknown function belongs to the set of continuous, monotonic increasing/decreasing and globally concave/convex functions. We describe the isoquant function $\mathcal{H}$ at $y^{(i)}$ as

$$X_{d,j}^{(i)} = \mathcal{H}\left(\boldsymbol{X}_{-d,j}^{(i)}; y^{(i)}\right) + e_j = \alpha_j^{(i)} + \boldsymbol{\beta}_j^{(i)\prime}\boldsymbol{X}_{-d,j}^{(i)} + e_j, \quad \forall j = 1, \ldots, n_i. \tag{B.4}$$

where e_j is the random error satisfying $E(e) = 0$, $\alpha_j^{(i)}$ and $\boldsymbol{\beta}_{\boldsymbol{j}}^{(i)}$ define the intercept and slope parameters that characterize the estimated set of hyperplanes.

We compute the CNLS estimator I times with $\{\boldsymbol{X}_{-d}^{(i)}, X_d^{(i)}\}_{i=1}^{I}$, and obtain the isoquant estimation $\hat{X}_d^{(i)} = \hat{\mathcal{H}}(\boldsymbol{X}_{-d,j}^{(i)}; y^{(i)}) = \hat{\alpha}_j^{(i)} + \hat{\boldsymbol{\beta}}_j^{(i)\prime}\boldsymbol{X}_{-d,j}^{(i)}$ at each isoquant level $y^{(i)}$. The CNLS estimator can be computed by solving the quadratic programming problem:

$$
\begin{aligned}
\min_{\alpha,\boldsymbol{\beta}} \quad & \sum_{j=1}^{n_i}\left(X_{d,j}^{(i)}-\left(\alpha_j^{(i)}+\boldsymbol{\beta}_j^{(i)\prime}\boldsymbol{X}_{-d,j}^{(i)}\right)\right)^2 \\
\text{subject to} \quad & \alpha_j^{(i)}+\boldsymbol{\beta}_j^{(i)\prime}\boldsymbol{X}_{-d,j}^{(i)} \geq \alpha_l^{(i)}+\boldsymbol{\beta}_l^{(i)\prime}\boldsymbol{X}_{-d,j}^{(i)}, \quad \forall j,l=1,\ldots,n_i \\
& \boldsymbol{\beta}_j^{(i)} \leq 0, \qquad\qquad\qquad\qquad\qquad\quad \forall j=1,\ldots,n_i
\end{aligned}
\tag{B.5}
$$

The first inequality constraints in (B.5) can be interpreted as a system of Afriat inequalities (Afriat (1972); Varian (1984)) to impose convexity. The second inequality constraints impose monotone decreasing. We note that the functional estimates resulting from (B.5) is unique only for the observed data points. Seijo and Sen (2011) and Lim and Glynn (2012) proved the consistency of the CNLS estimator. Also Chen and Wellner (2016) proves that the CNLS estimator attains $n^{-1/2}$ pointwise rate of convergence if the true function is piece-wise linear.

B.2.2.2 Directional CNLS

The CNLS estimator in the previous section assumes that the input data contains errors only in the d-th input direction while all input variables are typically measured with error. Kuosmanen and Johnson (2017) introduces the CNLS estimator within the directional distance function (DDF) framework. The DDF indicates the distance from a given sample vector to the estimated function in some pre-assigned direction. In our isoquant estimation, we can write the DDF function as follows:

$$
\vec{D}(\boldsymbol{X}_{-d,j}^{(i)}, X_{d,j}^{(i)}, \boldsymbol{g}^{X_{-d}}, g^{X_d}) = e_j, \quad \forall j=1,\ldots,n_i \tag{B.6}
$$

where $(\boldsymbol{g}^{X_{-d}}, g^{X_d}) \in \mathbb{R}^d$ is the pre-assigned error direction. We can choose the error direction $(\boldsymbol{g}^{X_{-d}}, g^{X_d})$ empirically from the density of the input data so that the chosen error direction can explain the noise in the input data well. Note that we need to normalize input data $\{\boldsymbol{X}_{-d,j}^{(i)}, X_d^{(i)}\}_{i=1}^{I}$ to have unit variance. We divide the inputs by their standard deviation. Normalizing inputs avoids that one input, measured on a large scale, will dominate other inputs, measured on a smaller scale.

Similar to the CNLS estimator, we compute the directional CNLS estimator I times with $\{\boldsymbol{X}_{-d,j}^{(i)}, X_d^{(i)}\}_{i=1}^{I}$, and obtain the isoquant estimation at each isoquant level $y^{(i)}$. The directional CNLS estimator can be computed by solving the quadratic programming problem:

$$
\begin{aligned}
\min_{\alpha,\boldsymbol{\beta},\gamma} \quad & \sum_{j=1}^{n_i} \left(\gamma_j^{(i)} X_{d,j}^{(i)} - \left(\alpha_j^{(i)} + \boldsymbol{\beta}_j^{(i)\prime} \boldsymbol{X}_{-d,j}^{(i)}\right)\right)^2 \\
\text{subject to} \quad & \alpha_j^{(i)} + \boldsymbol{\beta}_j^{(i)\prime} \boldsymbol{X}_{-d,j}^{(i)} - \gamma_j^{(i)} X_{d,j}^{(i)} \geq \alpha_l^{(i)} + \boldsymbol{\beta}_l^{(i)\prime} \boldsymbol{X}_{-d,j}^{(i)} - \gamma_l^{(i)} X_{d,j}^{(i)}, \quad \forall j, l = 1, \ldots, n_i \\
& \boldsymbol{\beta}_j^{(i)} \leq 0, \qquad \forall j = 1, \ldots, n_i \\
& \gamma_j^{(i)} \geq 0, \qquad \forall j = 1, \ldots, n_i \\
& \gamma_j^{(i)} g^{X_d} + \boldsymbol{\beta}_j^{(i)\prime} \boldsymbol{g}^{X_{-d}} = 1, \qquad \forall j = 1, \ldots, n_i
\end{aligned} \tag{B.7}
$$

This formulation introduces new coefficients $\gamma_j^{(i)}$ that represents marginal effects of the d-th input $X_d^{(i)}$ to the DDF. Similar to the CNLS estimator (B.5), first three constraints impose convexity and monotonicity in all input directions respectively. The last constraints are normalization constraints that ensure the translation property (Chambers et al. (1998)).

B.2.2.3 Averaging directional CNLS

The directional CNLS estimator in previous section assumes that the input data contains errors in potentially all variables, but in fixed ratios such that the over all error direction is $(\boldsymbol{g}^{X_{-d}}, g^{X_d})$. However, in observed production data, the errors in different components of the input vector, $\boldsymbol{X}_j^{(i)}$, may vary in length randomly. Particularly when estimating input isoquants, observations can be projected to the function orthogonally as shown in Figure B.1. Noise here is mainly caused by the projection of observations to particular isoquant level $y^{(i)}$. This issue will be further addressed in Section 3.3.3.

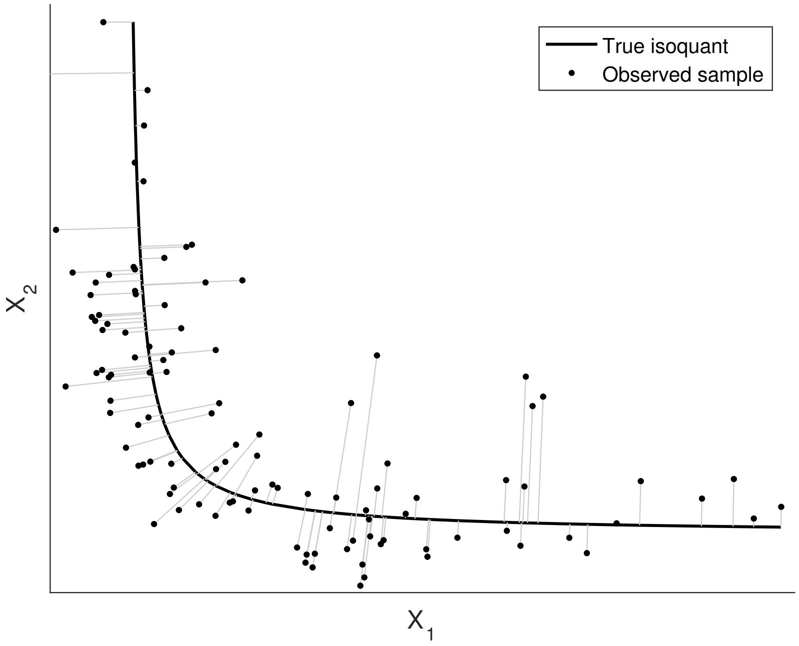


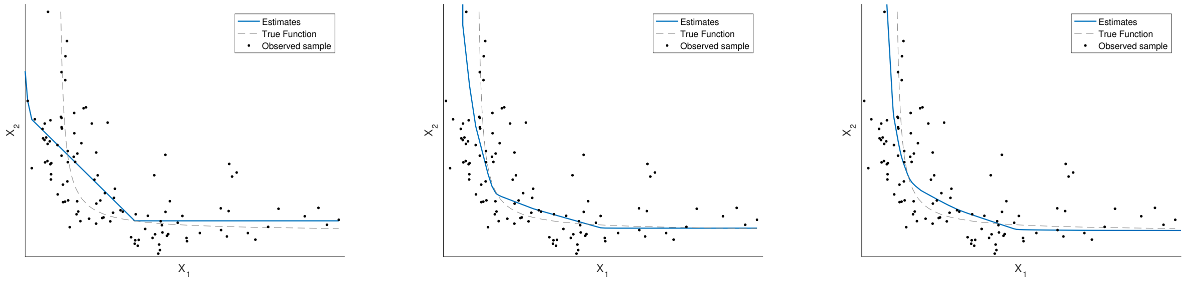
Figure B.1. Noise which is orthogonal to the true isoquant

If we misspecified the error direction, the estimated isoquants will be biased, and the bias will increase as the specified error direction is further from the true error direction. We propose a simple algorithm to average out a bias from the misspecification of the error direction. We define the set of error directions $\left\{\left(\boldsymbol{g}_m^{X_{-d}}, g_m^{X_d}\right)\right\}_{m=1}^{M}$ from the distribution of the input data $\boldsymbol{X}^{(i)}$ where M is the number of error directions considered.[1] For each isoquant level $y^{(i)}$, we compute the directional CNLS estimator (B.7) with each error direction $\left\{\left(\boldsymbol{g}_m^{X_{-d}}, g_m^{X_d}\right)\right\}_{m=1}^{M}$, and averaging them to obtain

[1] Based on our numerical experiments, we recommend to use $M = 10$ and define error directions by the equally spaced percentile of the input ratio.

the final isoquant estimates. The final isoquant estimates still satisfied conditions for an isoquant in Assumption 3.2 since the average of convex monotone decreasing functions is a convex monotone decreasing function.

Figure B.2 (a), (b) and (c) show the estimation results with CNLS, Direction CNLS and Averaging direction CNLS respectively with samples generated by radial errors. The CNLS estimator has significant bias for the observations for which noise is not only in X_1. Directional CNLS and averaging multiple estimates of directional CNLS with different directions performs better than the CNLS estimator because these methods allow for errors in all input dimensions. Although both extensions of CNLS do not result in a correctly specified the error direction for each observation, the methods perform well even for small sample size as shown in Appendix B.3.



(a) CNLS

(b) Directional CNLS

(c) Averaging directional CNLS

Figure B.2. Estimated isoquant by CNLS, Directional CNLS and Averaging directional CNLS

B.2.3 S-shape function

We are interested in estimating the S-shape function on rays from the origin as a component of our estimation procedure. This step is composed of two sub-steps: First, we project each observation to each ray from the origin by projecting along an estimated isoquant. Second, we estimate the S-shape function on each ray from the origin. We describe the procedure how to obtain the rays from the origin $\boldsymbol{\theta}^{(r)}$ in section 3.3.3.

B.2.3.1 Projecting an observation using the estimated isoquant information

Before estimating S-shape functions, we project the observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$ to each ray from the origin $\boldsymbol{\theta}^{(r)}$. We use the estimated isoquants described in Section B.2.2 to project the observations. First, for each observation, we extract the two estimated isoquants which sandwich the observation in input space. Figure B.3(b) shows the example that two isoquants sandwiching the observation $\boldsymbol{X}_j$.

Second, we compute the intersection of extracted isoquants and the ray from the origin to the observation, and define distances to the isoquants below and above as $r_j^{(below)}$ and $r_j^{(above)}$ respectively. Then we can compute the weights ρ_j which is defined as

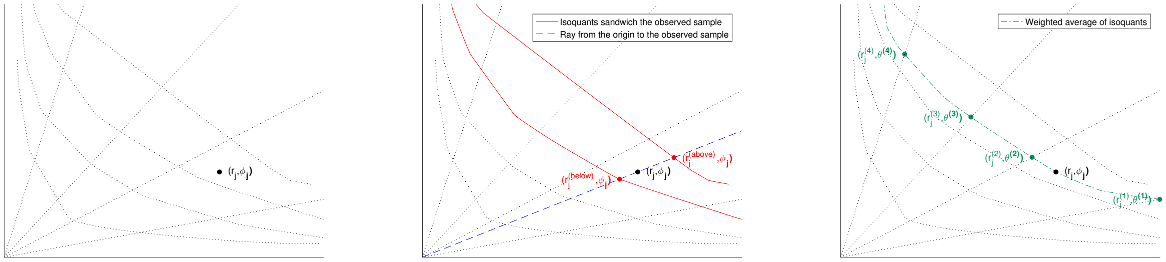
$$\rho_j = \frac{r_j - r_j^{(below)}}{r_j^{(above)} - r_j^{(below)}},\ j = 1, \dots, n \tag{B.8}$$

where $0 \leq \rho_j \leq 1$, and ρ_j approaches 1 as r_j is closer to $r_j^{(above)}$. Intuitively, we aim to use more information from the isoquant above when the observation is closer to the isoquant above. Figure B.3(b) also shows the definition of $r_j^{(below)}$ and $r_j^{(above)}$. In case that the observation is below or above the minimum or maximum isoquant, we define $r_j^{(below)} = 0$ and $r_j^{(above)} = r_j$ respectively.

Finally, we compute the intersection of extracted isoquants and each ray from the origin, and define distances to the intersection with isoquants below and above as $r_j^{(below)(r)}$ and $r_j^{(above)(r)}$ respectively for $r = 1, \dots, R$. Then we obtain the projected observation $\{r_j^{(r)}, \boldsymbol{\theta}^{(r)}\}_{r=1}^R$ as follows:

$$r_j^{(r)} = \left(r_j^{(above)(r)} - r_j^{(below)(r)}\right)\rho_j \quad \forall r = 1, \dots, R. \tag{B.9}$$

Figure B.3(c) shows the example of projection to each ray from the origin $\boldsymbol{\theta}^{(r)}$. Intuitively, we compute the inverse distance weighted average of two isoquants which sandwich the observation.



(a) Estimated isoquants and predefined rays from the origin

(b) Two isoquants sandwich the observation

(c) Projection of observation to each predefined rays from the origin

Figure B.3. Procedures of the projection of the observation in input space

B.2.3.2 Shape Constrained Kernel Least Squares (SCKLS)

Yagi et al. (2018) proposed the Shape Constrained Kernel-weighted Least Squares (SCKLS) which is a kernel-based nonparametric shape constrained estimator. The SCKLS estimator is an extension of Local Polynomial estimator (Stone (1977) and Cleveland (1979)) which imposes some constraints on parameters which characterize the estimated function such as intercept and slope. The SCKLS estimator introduces a set of G evaluation points to impose shape constraints on each evaluation point. We are now interested in estimating S-shape function on each ray from the origin. Define the evaluation points on a ray from the origin $\boldsymbol{\theta}^{(r)}$ as follows

$$r_g^{(r)} \in \{r_1^{(r)}, \dots, r_G^{(r)}\} \quad \forall r = 1, \dots, R. \tag{B.10}$$

Note that evaluation points in input space on ray r are defined by the scalar value $r_g^{(r)}$ which is a distance from the origin on the r^{th} ray.

The objective function of the SCKLS estimator uses kernel weights, so more weight is given to the observations that are closer to the evaluation point. In our S-shape estimation, there exist two different weights to be considered: 1) the angle between the observation and the ray from the origin for which we are currently estimating, 2) the distance measured along the ray after the sample is projected using the estimated isoquant. Figure B.4 shows two different kernel weights imposed in our S-shape estimator.

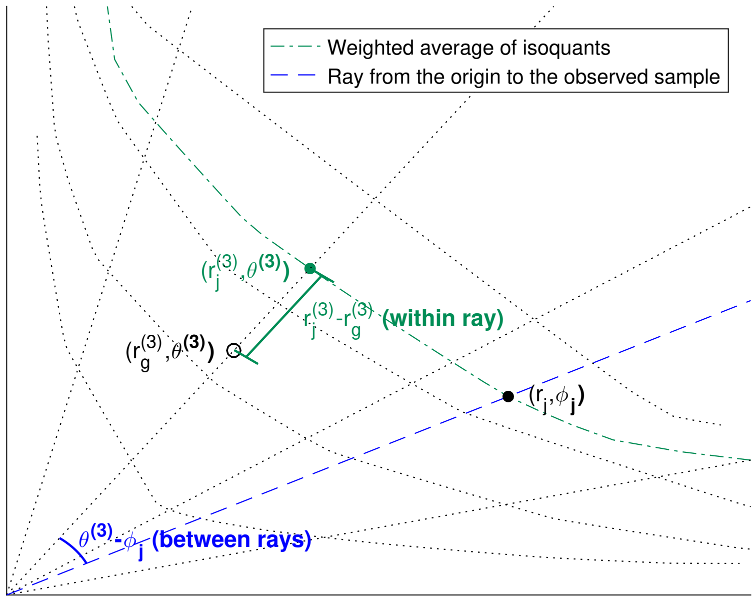

Weighted average of isoquants
Ray from the origin to the observed sample
$(r_j^{(3)}, \theta^{(3)})$
$r_j^{(3)} - r_g^{(3)}$ (within ray)
$(r_g^{(3)}, \theta^{(3)})$
(r_j, ϕ_j)
$\theta^{(3)} - \phi_j$ (between rays)

Figure B.4. Kernel weight in the S-shape estimation

Here, we define a distance measure in angles by the relative distance to the neighboring rays in each ray dimension $k = 1, \ldots, d-1$

$$D\left(\phi_{jk}, \theta_k^{(r)}\right) = \begin{cases} \frac{\phi_{jk} - \theta_k^{(r)}}{\theta_k^{(r+1)} - \theta_k^{(r)}} & \phi_{jk} \geq \theta_k^{(r)} \\ \frac{\theta_k^{(r)} - \phi_j}{\theta_k^{(r)} - \theta_k^{(r-1)}} & otherwise \end{cases} \tag{B.11}$$

Intuitively, the distance is greater than one if the observation is outside of the cone defined by the adjacent rays.

For each ray from the origin $\boldsymbol{\theta}^{(r)}$, we solve the following mixed-integer quadratic programming problem:

$$\begin{aligned} \min_{\boldsymbol{a}, \boldsymbol{b}, g_*^{(r)}} \quad & \sum_{g=1}^{G} \sum_{j=1}^{n} \left(y_j - \left(a_g^{(r)} + b_g^{(r)} \left(r_j^{(r)} - r_g^{(r)} \right) \right) \right)^2 K \left(\frac{D\left(\boldsymbol{\phi}_j, \boldsymbol{\theta}^{(r)}\right)}{\boldsymbol{\omega}} \right) k \left(\frac{r_j^{(r)} - r_g^{(r)}}{h^{(r)}} \right) \\ \text{subject to} \quad & a_g^{(r)} - a_l^{(r)} \leq b_g^{(r)} \left(r_g^{(r)} - r_l^{(r)} \right) \quad \forall g, l = 1, \ldots, g_*^{(r)} - 1 \\ & a_g^{(r)} - a_l^{(r)} \geq b_g^{(r)} \left(r_g^{(r)} - r_l^{(r)} \right) \quad \forall g, l = g_*^{(r)}, \ldots, G \\ & b_g^{(r)} \geq 0 \quad \forall g, l = 1, \ldots, G \end{aligned} \tag{B.12}$$

where $a_g^{(r)}$ is a functional estimate, $b_g^{(r)}$ is an estimate of the slope of the function at $r_g^{(r)}$, the g^{th} evaluation point on the r^{th} ray. $k(\cdot)$ and $K(\cdot)$ denote the kernel and the product kernel function respectively. The observation which is closer to the evaluation points as measured by the angular deviation, and along the projected ray gets more weight in the estimation. $\boldsymbol{\omega}$ and $h^{(r)}$ are tuning parameters for the kernel estimator which we will refer to as bandwidths. The first and second constraints in (B.12) are the convexity and concavity constraints respectively. We also need to estimate an index of an inflection point $g_*^{(r)}$ which is the point at which the S-shape function switches from convex to concave. We solve the quadratic programming problem G-times, once for each value $g_*^{(r)} \in \{1, \ldots, G\}$, and obtain a S-shape estimation by selecting the solution which has the minimum objective value among these G solutions.

B.2.3.3 Update isoquant estimates

After estimating the S-shape function along each ray, we need to verify whether the estimated S-shape functions satisfy the input convexity assumption. For this purpose, we cut the S-shape estimates at each isoquant level, $\boldsymbol{y}^{(i)}$, and obtain intersecting points defined by radial coordinates as $\{r^{(i)(r)}, \boldsymbol{\theta}^{(r)}\}$. Figure B.5 shows how we obtain the intersecting points $\{r^{(i)(r)}, \boldsymbol{\theta}^{(r)}\}$ with 2-input example. We now re–estimate the isoquants by applying the CNLS-based method to the intersections for each isoquant $\{r^{(i)(r)}, \boldsymbol{\theta}^{(r)}\}_{r=1}^{R}$. Note that we can convert this into Cartesian coordinate system through the inverse of the equations shown in (3.5), and apply the CNLS-based method explained in Appendix B.2.2.

B.2.3.4 Minimizing the gap between estimates

We now have computed both S-shape and isoquant estimates. If the S-shape estimates do not violate the input convexity assumption, then the functional estimates of the S-shape functions and the input isoquants should match at each isoquant y-level. However if the S-shape estimates violate the input convexity assumption, then the S-shape estimates will not match the isoquant estimates at some isoquant y-level as shown in Figure B.6(a) with a blue circle. Here we propose to solve a quadratic programming problem which aims to minimize the gap between S-shape and isoquant estimates.

In this problem, we try to modify the S-shape estimates while fixing an inflection point at the same position as the original S-shape estimates. The objective function computes the weighted average of two deviations: 1) a gap between original S-shape estimates and revised S-shape esti-

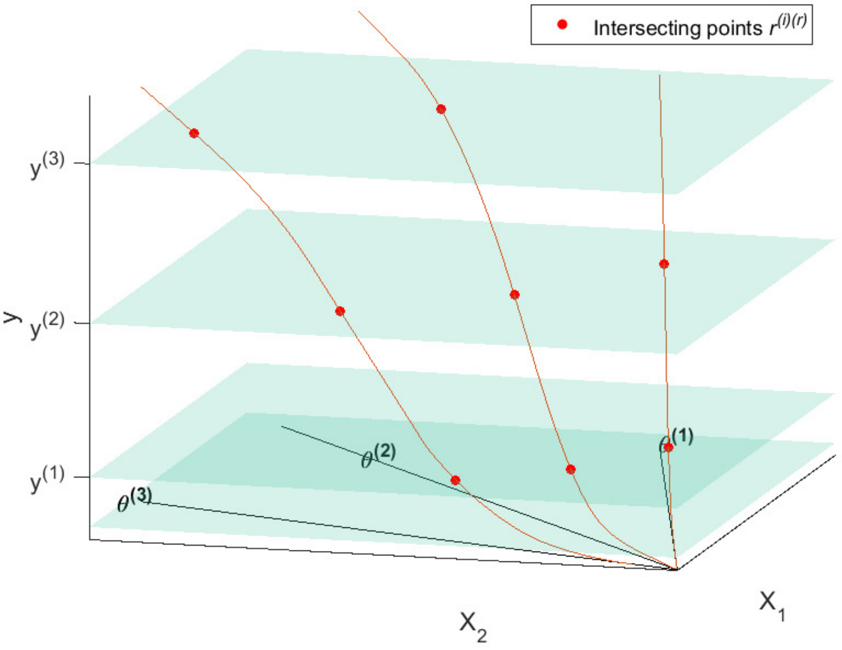
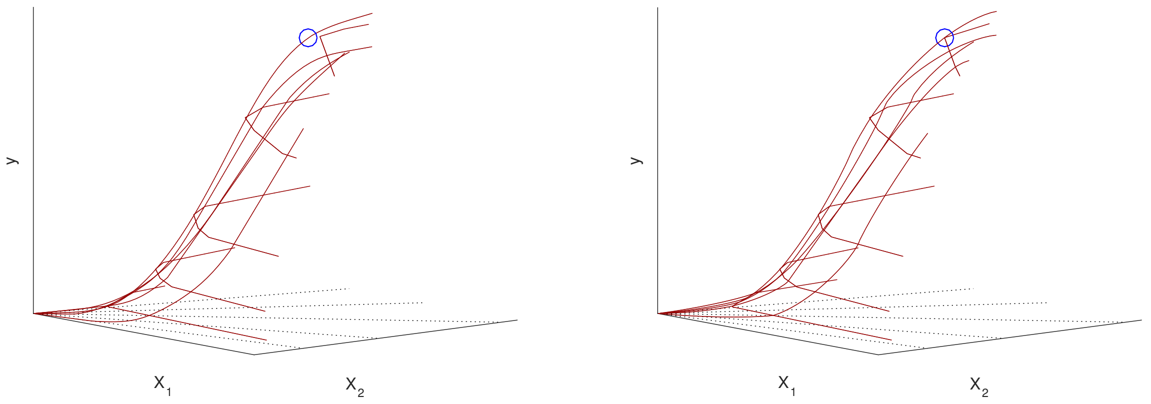


Figure B.5. How to obtain intersecting points $r^{(i)(r)}$

mates, and 2) a gap between revised S-shape estimates and the isoquant estimates at each isoquant y-level. Intuitively, we want to obtain the revised S-shape estimates which is close to the original S-shape estimates while satisfying input convexity. Figure B.6(b) shows the example that a violation is resolved through this step.

Here, we describe the mathematical formulation. We start from redefining the evaluation points on a ray, $\boldsymbol{\theta}^{(r)}$ as

$$\begin{aligned} & r_g^{(r)} \in \{r_1^r, \ldots, r_G^r\} && \forall g = 1, \ldots, G \\ & r_{g^{(i)}}^{(r)} \in \{r^{(1)(r)}, \ldots, r^{(I)(r)}\} && \forall i = 1, \ldots, I \\ & r_{g'}^{(r)} \in \{r_1^r, \ldots, r_G^r\} \cup \{r^{(1)(r)}, \ldots, r^{(I)(r)}\} && \forall g' = 1, \ldots, G' \end{aligned} \tag{B.13}$$



(a) Before modification

(b) After modification

Figure B.6. Modification of S-shape estimates

where $G' = G + I$. $r^{(r)}_{g^{(i)}}$ is the intersecting points obtained in section B.2.3.3 and they are added to the set of evaluation points, $r^{(r)}_{g'}$. We aim to minimize the gap between S-shape and isoquant estimates by solving the following quadratic programming problem:

$$\begin{aligned}
\min_{\tilde{a}_g^{(r)}} \quad & w^S \cdot \frac{1}{R \cdot G} \sum_{r=1}^{R} \sum_{g=1}^{G} \left(\tilde{a}_g^{(r)} - a_g^{(r)}\right)^2 + w^I \cdot \frac{1}{R \cdot I} \sum_{r=1}^{R} \sum_{i=1}^{I} \left(\tilde{a}_{g^{(i)}}^{(r)} - y^{(i)}\right)^2 \\
\text{subject to} \quad & \frac{\tilde{a}_{g+2}^{(r)} - \tilde{a}_{g+1}^{(r)}}{r_{g+2}^{(r)} - r_{g+1}^{(r)}} \geq \frac{\tilde{a}_{g+1}^{(r)} - \tilde{a}_{g}^{(r)}}{r_{g+1}^{(r)} - r_{g}^{(r)}} \qquad \forall r \text{ and } \forall g = 1, \ldots, g_*^{(r)} - 2 \\
& \frac{\tilde{a}_{g+2}^{(r)} - \tilde{a}_{g+1}^{(r)}}{r_{g+2}^{(r)} - r_{g+1}^{(r)}} \leq \frac{\tilde{a}_{g+1}^{(r)} - \tilde{a}_{g}^{(r)}}{r_{g+1}^{(r)} - r_{g}^{(r)}} \qquad \forall r \text{ and } \forall g = g_*^{(r)} - 2, \ldots, G \\
& \tilde{a}_{g+1}^{(r)} \geq \tilde{a}_{g}^{(r)} \qquad\qquad \forall r \text{ and } \forall g = 1, \ldots, G
\end{aligned} \tag{B.14}$$

where $\tilde{a}_g^{(r)}$ denotes a revised functional estimate at a grid point g on a ray r. w^S and w^I are the weights for the S-shape estimator[2] and the isoquant estimator respectively satisfying $w^S, w^I \in$

[2] We set $w^S = 0.1$ and $w^I = 0.9$ for our simulation and application to make sure the gap between isoquants and S-shape estimates become small for every single iteration.

$[0, 1]$ and $w^S + w^I = 1$. The objective function computes the weighted average of two deviations: 1) a gap between original S-shape estimates and revised S-shape estimates, 2) a gap between revised S-shape estimates evaluated at the input vectors located on the estimated isoquant and isoquant level $y^{(i)}$. Intuitively, when we put more weight on the original S-shape estimate, w^S is large, the revised S-shape is close to the original S-shape, and input convexity may be violated. In contrast, when we put more weight on the isoquant estimates, w^I is large, the revised S-shape can be far from the original S-shape, but the resulting estimate is more likely to satisfy input convexity without any violations. Based on our numerical experiments, we recommend to set a larger value of w^I to avoid violations of the input convexity.

Constraints in (B.14) correspond to constraints in (B.12). First two constraints impose the convexity and concavity for the RUP law, and the last constraint imposes the estimated function is monotonically increasing.

B.2.3.5 Computing functional estimates on observations

The last step of an iteration is obtaining the functional estimates $\hat{g}(\boldsymbol{x})$ at any given value of input vector $\boldsymbol{x}$, and compute MSE against observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^n$. The procedure to compute a functional estimate on $\boldsymbol{x}$ is composed of two steps: 1) Compute the weighted average of the two closest isoquants to $\boldsymbol{x}$, and 2) Compute the weighted average of S-shape estimates on each ray. The first step is explained in appendix B.2.3.1. In this step, we obtain projected input data $\{r_j^{(r)}, \boldsymbol{\theta}^{(r)}\}_{r=1}^R$ which is defined in equation (B.9).

Then we can compute the functional estimates $\tilde{a}_j^{(r)}$ at projected ray $\{r_j^{(r)}, \boldsymbol{\theta}^{(r)}\}_{r=1}^R$ by linear interpolating revised S-shape estimates $\tilde{a}_g^{(r)}$ obtained in (B.14). Subsequently, we can compute the inverse distance weighted average of functional estimates by

$$\hat{g}(X_j) = \begin{cases} \tilde{a}_j^{(r)} & \exists\, r \text{ such that } d(\boldsymbol{\theta}_j, \boldsymbol{\theta}^{(r)}) = 0 \\ \frac{\sum_{r=1}^{R} p_j^{(r)} \tilde{a}_j^{(r)}}{\sum_{r=1}^{R} p_j^{(r)}} & otherwise \end{cases} \tag{B.15}$$

where $p_j^{(r)}$ is the inverse distance weight defined by

$$p_j^{(r)} = \frac{1}{d(\boldsymbol{\theta}_j, \boldsymbol{\theta}^{(r)})} \quad \forall r = 1, \ldots, R \tag{B.16}$$

where $d(\cdot)$ denotes a Euclidean distance function between two angles defined by

$$d(\boldsymbol{\theta}_j, \boldsymbol{\theta}^{(r)}) = \left\| \boldsymbol{\theta}_j - \boldsymbol{\theta}^{(r)} \right\| \tag{B.17}$$

Finally, we can compute the MSE against observations $\{\boldsymbol{X}_j, y_j\}_{j=1}^{n}$ as

$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \hat{g}(\boldsymbol{X}_j) \right)^2 . \tag{B.18}$$

B.2.4 Updating parameters

Finally, we update the parameters for the estimation before moving forward to the next iteration. We first update the parameters defining the number of both isoquants and rays to be estimated. When the gap between isoquants and S-shape estimates is large for a certain consecutive iterations, we delete the corresponding isoquant or ray. Specifically for any ray r, if $\left(\frac{y^{(i)} - \hat{g}(\boldsymbol{X}^{(i)(r)})}{y^{(i)}} \right) > \delta$ for some isoquant i for C consecutive iterations, then delete ray r where δ is a tolerance value of percentage errors and C is a number of consecutive iterations allowing errors over the tolerance.[3][4] And similarly defined for isoquant i.

We also update the bandwidth between rays, $\boldsymbol{\omega}$, used in the SCKLS-based S-shape estimation. We update the value of $\boldsymbol{\omega}$ increasing it by $\Delta\boldsymbol{\omega}$ in each iteration. As an iteration goes forward, the bandwidth $\boldsymbol{\omega}$ becomes larger. We continue iterations until $\boldsymbol{\omega}$ becomes large enough that the functional estimates are stable between iterations and then we select the results of the iteration with the lowest MSE among the solutions with

$$\left(\frac{y^{(i)} - \hat{g}(\boldsymbol{X}^{(i)(r)})}{y^{(i)}}\right) \leq \delta \quad \forall r = 1, \ldots, R \text{ and } \forall i = 1, \ldots, I.$$

Since the algorithm start from a small value of $\boldsymbol{\omega}$, the S-shape function only uses observations close to the ray for the estimation. As the iterative algorithm proceeds, the S-shape estimator includes observations which are more distant from the ray on which the evaluation point under consideration lies. Thus, the estimated functions on each ray becomes more similar as the bandwidth increases. If there still exists a gap between S-shape and input isoquant estimates even with large $\boldsymbol{\omega}$, we delete the corresponding isoquant or ray following the rule described above. Thus, the gap between the S-shape estimates and the isoquant estimates can be made arbitrarily small by deleting isoquants. This characteristic of the algorithm will be used to prove the convergence of our iterative algorithm because a production function estimate with only one isoquant estimate is a homothetic production function and our estimation procedure has no gap for estimating functions that satisfying the RUP law and are homothetic in inputs.

[3] We allow large errors for $C = 10$ iterations for our simulation studies.

[4] We use $\delta = 0.01$ or $\delta = 0.05$ in our implementation depending on the noise size of data set.

B.3 Comparison of different input isoquant estimation methods

In section B.2.2, we introduce three different methods to estimate convex input isoquants: Convex Nonparametric Least Squares (CNLS), Directional Convex Nonparametric Least Squares (DCNLS) and Averaging Convex Nonparametric Least Squares (ADCNLS). In this section, we compare the performance of these estimators through Monte Carlo simulations.

We consider the following convex isoquant with 2-input.

$$X_2 = \mathcal{H}(X_1) = \frac{a}{X_1} \tag{B.19}$$

where a defines the shape of convex isoquant, and we use $a = 10$ in this experiment. Two-input satisfying equation B.19 is generated by

$$\begin{aligned} X_{1j}^* &= \sqrt{\frac{a}{\tan(\eta_j)}} \quad \forall j = 1, \ldots, n \\ X_{2j}^* &= \sqrt{a \cdot \tan(\eta_j)} \quad \forall j = 1, \ldots, n \end{aligned} \tag{B.20}$$

where angles η_j are randomly generated by $\eta_j \sim unif(0.05, \frac{\pi}{2} - 0.05)$. Then we generate samples by adding noise in the direction orthogonal to the true function.

$$\begin{aligned} X_{1j} &= X_{1j}^* + \epsilon_j \cdot \cos\left(\arctan\left({X_{1j}^*}^2/a\right)\right) \quad \forall j = 1, \ldots, n \\ X_{2j} &= X_{2j}^* + \epsilon_j \cdot \sin\left(\arctan\left({X_{1j}^*}^2/a\right)\right) \quad \forall j = 1, \ldots, n \end{aligned} \tag{B.21}$$

where additive noise ϵ_j is generated by $\epsilon_j \sim N(0, \sigma_v)$.

We consider 9 different scenarios with the different training sample size $n \in (50, 100, 200)$ and the standard deviation of the noise $\sigma_v \in (0.5, 1.0, 1.5)$. We use $M = 10$ different error directions for estimating ADCNLS where error directions are chosen by equally spaced percentiles of the input ratio $\{X_{2j}/X_{1j}\}_{j=1}^{n}$. We generate 100 training-testing set pairs for each scenario, and draw box plots of RMSE against the true function for each estimator on testing set in Figure B.7. Note that RMSE is computed in the direction orthogonal to the true function. The size of the testing set is 1000, and it is randomly drawn from the same distribution as the training set.

The DCNLS and ADCNLS estimators perform better than the CNLS estimator because these estimation methods assume errors are contained in both input dimensions. Although these two estimators still have misspecification of error directions, it helps to reduce the bias caused by the misspecification of error directions in the CNLS estimator.

B.4 Differences Between S-shape Definition and the RUP Law

In this section, we provide an example in which a production function that satisfies the RUP law, Definition 3.4, contains multiple inflection points.

Consider the following univariate example.

Example 1.

$$g(x) = x^{(1.8)} \exp(-x) \exp\left(\frac{-x \sin(100x)}{10000}\right)$$

Then we can compute the elasticity of scale and its derivative.

$$\epsilon(x) = 1.8 - x\{\frac{\cos(100x)}{100} + 1\},$$

Figure B.8 shows the elasticity of scale, $\epsilon(x)$, is monotonically decreasing on $x \in [0, 1]$ from 1.8 to 0.8, which satisfies Definition 3.4. Figure B.9 shows that the production function and its first and second derivative respectively. In Figure B.9 (a), the production function looks S-shape;

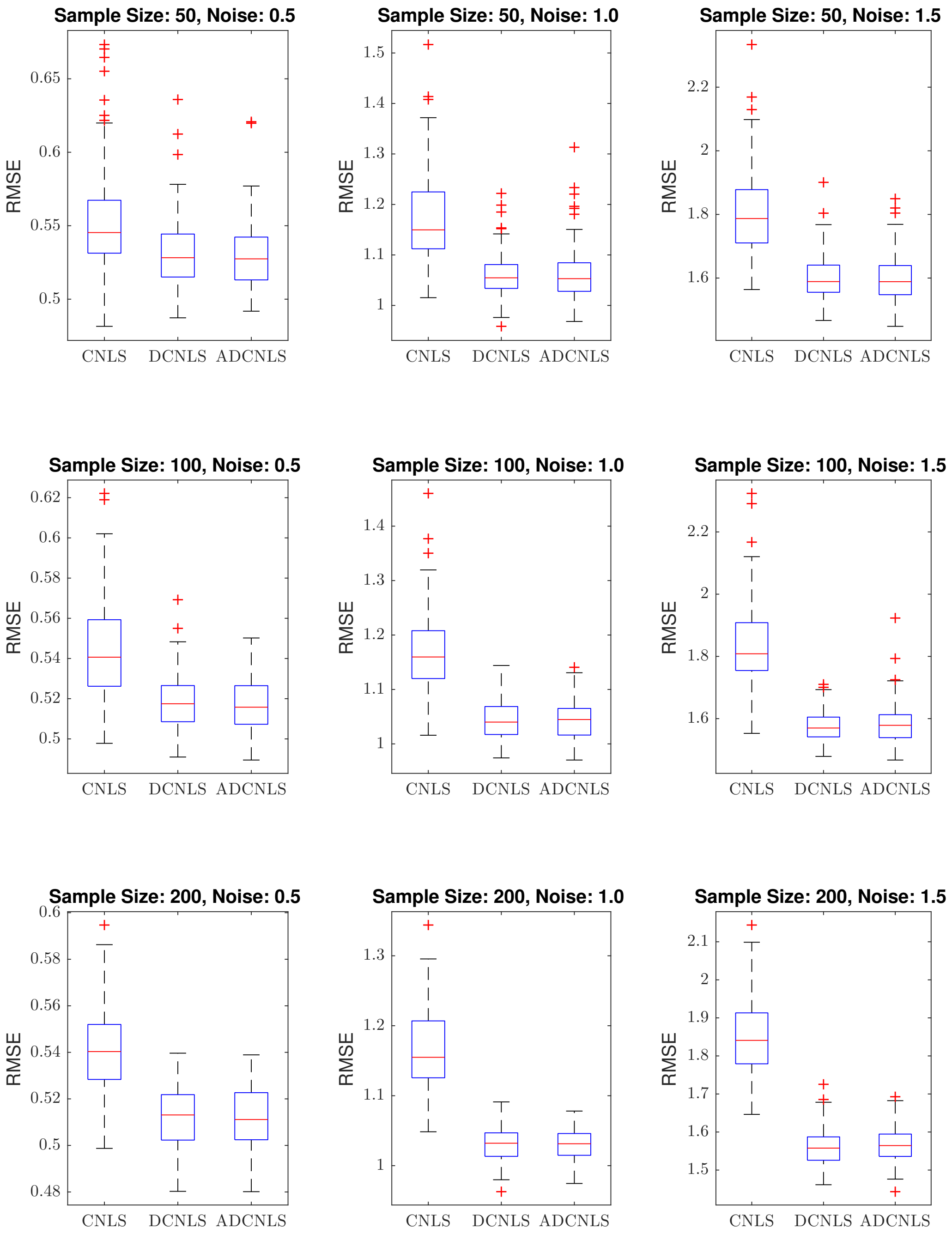


Figure B.7. Estimation results on the testing set for the isoquant estimation

however, Figure B.9 (c) shows that the production function has a multiple inflection points as there are multiple intersections between its second derivative $g''(x)$ and constant function at $x = 0$. So this is a counterexample of S-shape with the RUP law. Thus, to avoid having multiple inflection points, we added the condition on the second derivative of the function $g_0(\cdot)$ as shown in Definition 3.5.

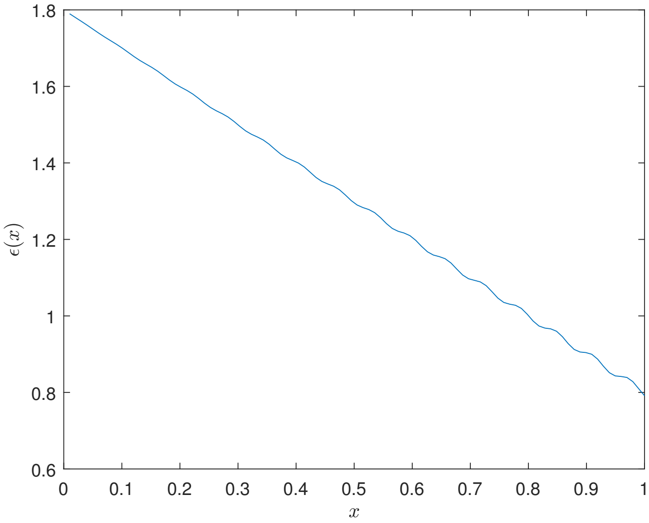
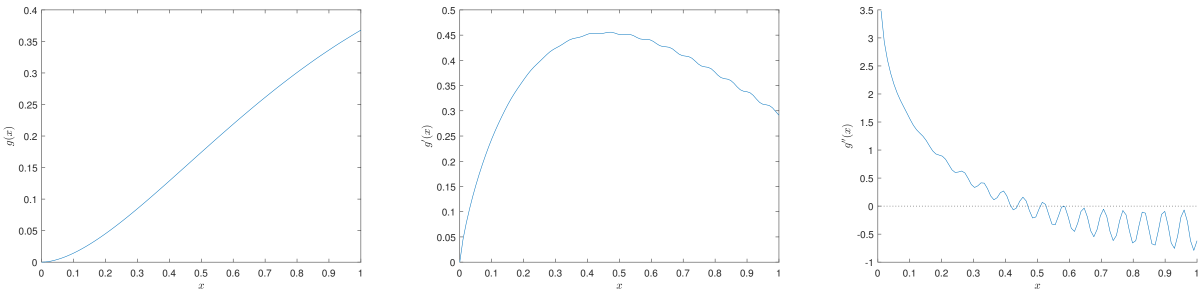


Figure B.8. The elasticity of scale



(a) Production function

(b) First derivative

(c) Second derivative

Figure B.9. Production function and its derivatives

APPENDIX C

APPENDIX OF CHAPTER 4

C.1 Comprehensive results of the application to the Japanese homogeneous products industries

We show the comprehensive summary of results in Section 4.4. First, we clarify the trend of each industry by displaying the transition of input and value added from 1997 to 2007. In the main manuscript, we show Figure 4.5 describing the transition of the full–time and part–time labor headcount. Here, Figures C.1 and C.2 show the transition of capital input and value added respectively. These figures describe the expansion and shrinkage of each industry. Specifically, sugar, bread, coffee and cardboard industries seem stable over years while plywood, ready–mix concrete and concrete products industries are shrinking. Products belonging to the shrinking industries are used for housing, construction or infrastructure whose demand was declined from 1997 to 2007 due to the Asian financial crisis and decline of youth population.

Figure C.3 shows the transition of aggregated productivity level by Cobb–Douglas OLS, Cobb–Douglas 2SLS, S–shape without IV and S–shape with IV respectively. This is the comprehensive results of Figure 4.7. Figure C.4 shows the estimated MPSS by S–shape with IV and S–shape without IV models. Sugar and coffee industries have relatively noisy estimates. This is caused by the small sample size since few observations may change the shape of production function and MPSS estimates drastically. For bread industry, S–shape with IV and without IV models have significantly different MPSS estimates. As shown in Figure 4.5, firms in bread industry tend to adjust their part–time labors, and thus, total labor headcount is likely to be endogenous. Thus, S–shape without IV model has inaccurate MPSS estimates due to the bias from the endogeneity. Another potential cause is that due to the lack of the variation in input ratio of observations, our model may fail to obtain the reasonable isoquant estimates. This may be also the cause of the noisy MPSS estimates in cardboard industry. One of the potential solutions is to select rays and isoquant y-levels carefully to make sure each estimates can capture the characteristics in the observations. For example, we can use K-means clustering to define the group of observations, and define the centroid of each group as a ray and isoquant y–level. Plywood, ready–mix and concrete products have relatively similar estimates of MPSS although there are some deviations.

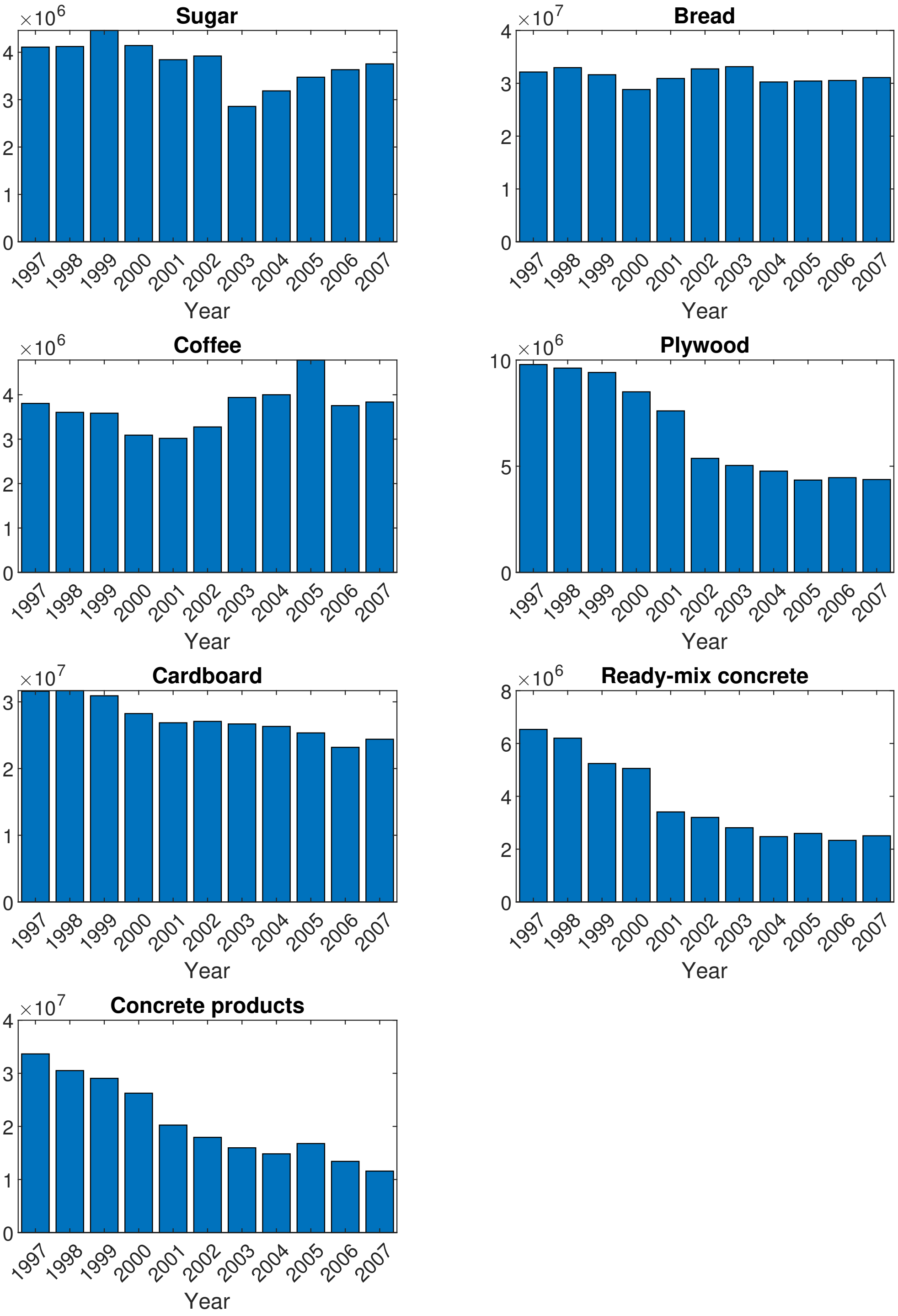


Figure C.1. Transition of capital input

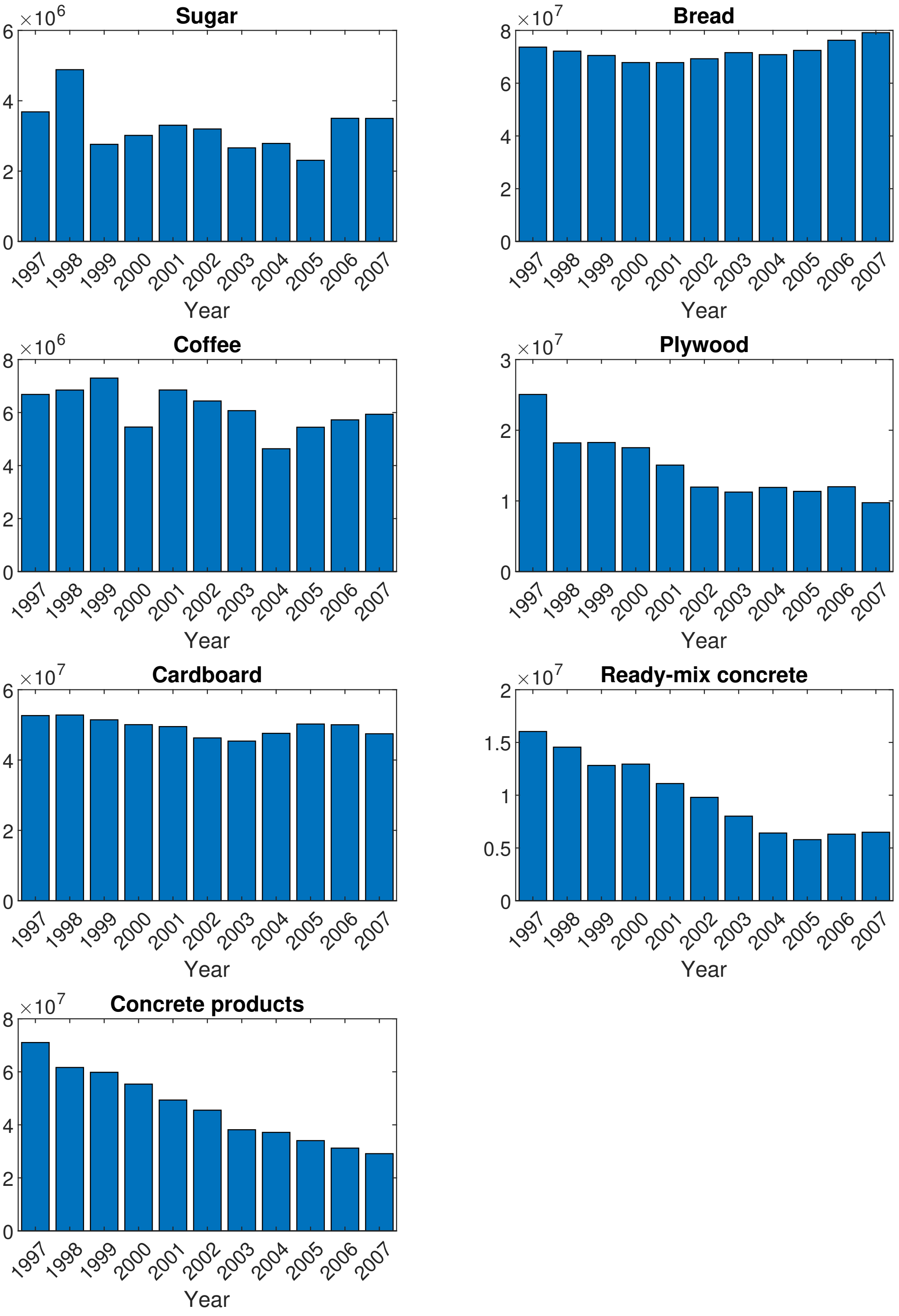


Figure C.2. Transition of value added

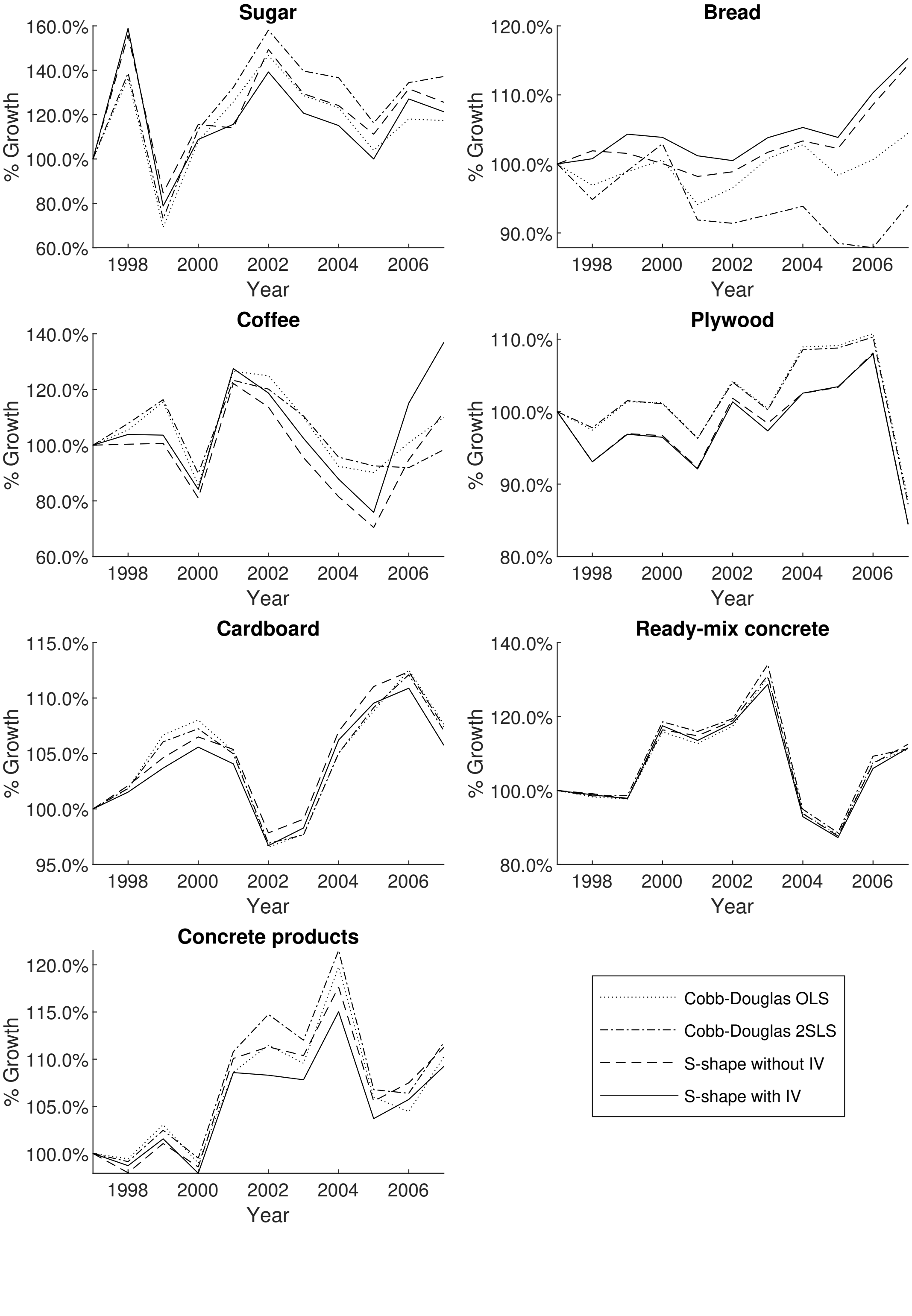


Figure C.3. Percentage growth of the aggregated productivity from 1997

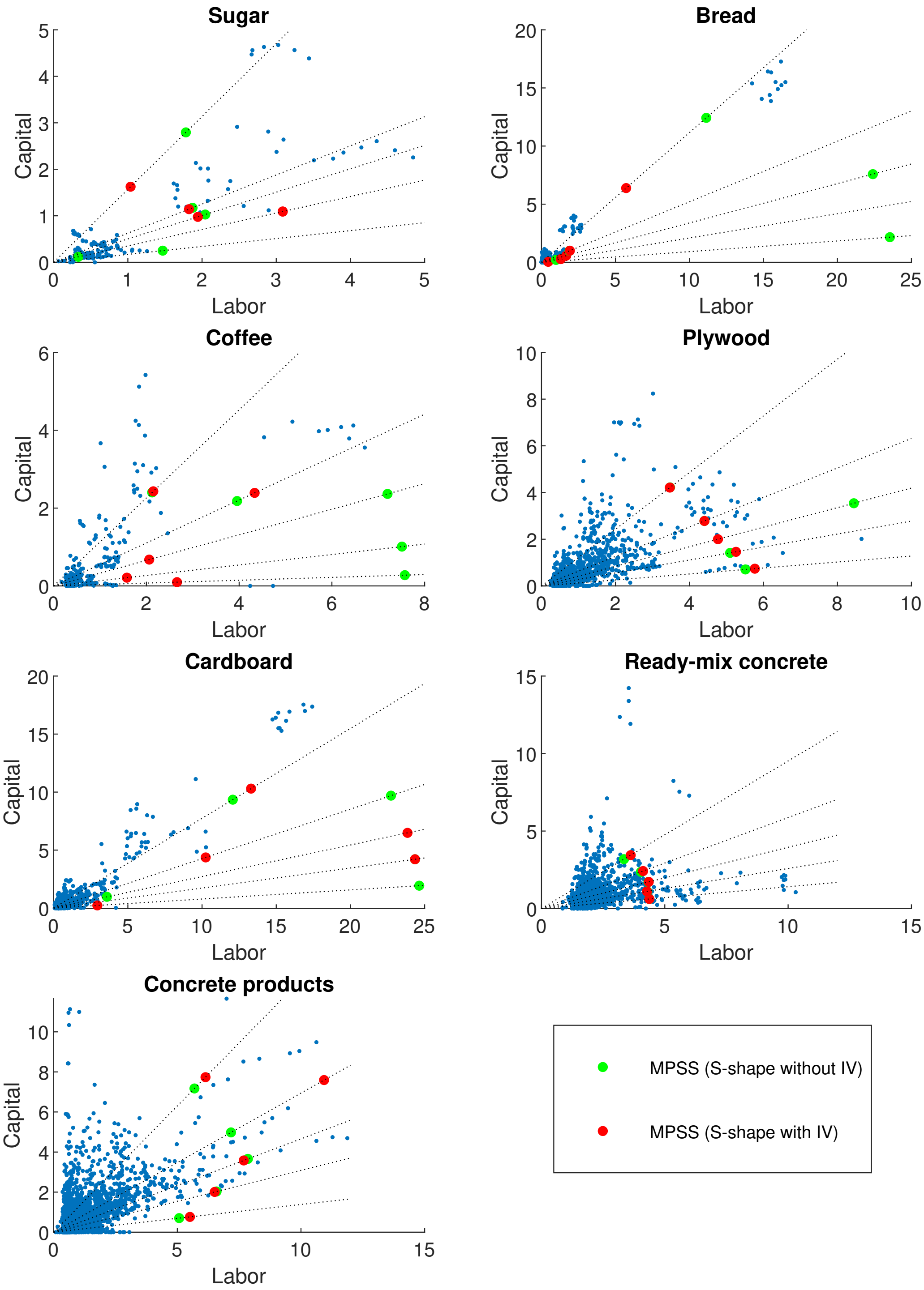


Figure C.4. Estimated MPSS by S–shape with IV and S–shape without IV