

NONPARAMETRIC ESTIMATION OF CONCAVE PRODUCTION TECHNOLOGIES BY ENTROPIC METHODS

GAD ALLON,^a MICHAEL BEENSTOCK,^{b*} STEVEN HACKMAN,^c URY PASSY^d AND
ALEXANDER SHAPIRO^c

^a *Kellogg School of Management, Northwestern University, USA*

^b *Department of Economics, Hebrew University of Jerusalem, Jerusalem, Israel*

^c *School of Industrial and Systems Engineering, Georgia Institute of Technology Atlanta, GA USA*

^d *Faculty of Industrial Engineering and Management Technion—Israel Institute of Technology, Haifa, Israel*

SUMMARY

An econometric methodology is developed for nonparametric estimation of concave production technologies. The methodology, based on the principle of maximum likelihood, uses entropic distance and convex programming techniques to estimate production functions. Empirical applications are presented to demonstrate the feasibility of the methodology in small and large datasets. Copyright © 2007 John Wiley & Sons, Ltd.

Received 14 September 2004; Revised 4 October 2005

1. INTRODUCTION

The econometric analysis of production functions has a long history, dating back to the pioneering efforts of Cobb and Douglas (1928). A constant theme to this history has been the search for ever more flexible functional forms. The legendary Cobb–Douglas production function assumes that the elasticity of substitution (ES) between factors of production is unity and returns to scale (RTS) are constant. Arrow *et al.* (1961) relaxed the restriction that $ES = 1$ in their CES production function, but assumed that ES is constant. Subsequently, Christensen *et al.* (1973) proposed the translog production function (TL) by permitting ES to vary between different factors of production and at different scales of output. Lau (1986) provides a survey of these, and related, developments. More recently, Zellner and Ryu (1998) suggest using the Box–Cox transformation to allow RTS and ES to vary. Below we follow Zellner and Ryu and refer to this highly flexible functional form by NRVES.¹ A further advantage of NRVES is that, unlike TL, it is quasiconcave and therefore satisfies the neoclassical properties of a production function.

In this paper we suggest a nonparametric methodology for estimating production functions. We make no parametric assumptions about the distribution of the disturbances, and only the weakest of assumptions about functional form. We assume the production function is non-negative, nondecreasing, and concave (diminishing marginal returns).² We show that the maximum likelihood (ML) estimation problem may be equivalently formulated as a convex program. For large-size problems, where either or both the number of observations and the number of factors

* Correspondence to: Michael Beenstock, Department of Economics, Hebrew University of Jerusalem, Jerusalem 91905, Israel. E-mail: msbin@mscc.huji.ac.il

¹ NR refers to Nerlove and Ringstad, who suggested the Box–Cox specification, and VES refers to variable ES.

² This echoes Manski (1995, Ch. 7), who makes the minimal assumption that the demand curve slopes downwards for purposes of nonparametric estimation of demand schedules.

of production are large, we explain how one may approximate the convex program with a *linear* program. Our approach is therefore feasible when there are several factor inputs and hundreds of data points.

Our approach is based upon the theoretical work of Hanoch and Rothschild (1972) and Afriat (1971, 1972), who suggested a sort of litmus test for quasiconcavity, monotonicity, and homotheticity given empirical data on inputs, outputs, and prices, or inputs and outputs only. The basic idea dates back to Afriat (1967) in the context of consumer spending. Hanoch and Rothschild's idea was to check the data to see whether the isoquants happen to cross each other or bend in the 'wrong' direction. They clearly saw the possibility of turning their approach into a production function estimator, but they desisted because they were reluctant to make 'blithe parametric assumptions' about the disturbances. They were also concerned about the computational complexities involved, especially when there are several factors of production and the number of observations is large.

In a series of papers Varian extended Hanoch and Rothschild's and Afriat's idea to consumer data (Varian, 1982, 1983) and to production data (Varian, 1984, 1985).³ He asked if there exists a well-behaved production function that is empirically consistent with cost minimization or profit maximization. Our efforts are similar in spirit to Varian's. However, they differ in several important respects. First, we show that our estimator is consistent. Secondly, we use convex programming rather than quadratic programming to carry out the optimization. Third, we demonstrate that our methodology works even when the sample size is large and when the number of factors of production exceeds 2, as in Varian's example. Fourth, unlike Varian, we do not have price data, so we test for concavity and homotheticity rather than cost minimization. The existence of price data naturally provides more information, thereby easing the burden of estimation. Therefore our estimation problem is more challenging than Varian's.

Banker and Maindiratta (1992) suggested a similar idea to ours, except they decompose the noise into two components: measurement error, which is assumed to be normally and identically distributed; and optimization (efficiency) error, which is assumed to be truncated (at zero) normal. Like us, they assume that output varies directly with inputs and that the technology is convex. Because Banker and Maindiratta did not apply their methodology to empirical data, it is difficult to judge whether it is feasible.⁴ By contrast, we demonstrate with empirical examples that our methodology is feasible, even when the sample size is large, and we do not make arbitrary parametric assumptions about the noise.

Our proposed estimator joins a small but expanding literature on nonparametric estimation subject to shape constraints. The two key shape constraints under consideration here are monotonicity and concavity. Statisticians, e.g. Hall and Huang (2001), have devoted considerable attention to the former but not the latter. Econometricians, however, have focused upon both monotonicity and concavity. Zellner and Ryu (1998) suggest a semiparametric procedure in which both monotonicity and concavity apply. Yatchew and Bos (1997) use penalized least squares to estimate monotonic and concave functions. Finally, Matzkin (1999) develops a specialized algorithm for nonparametric estimation of concave technologies under a variety of general shape constraints, and demonstrates it by solving a number of test problems involving two inputs and approximately

³ Matzkin (1991, 1993) considered the case where the variable of interest is discrete, as in consumer choice theory, rather than continuous, as here.

⁴ We are doubtful if it is feasible because their Problem 3 is a bi-level programming problem, and their Problem 4 is a nonconvex programming problem, both of which are very difficult to solve.

100 data points. The advantages of the convex and linear programming approach presented herein are twofold: first, convex and linear programming software are readily available; and second, our approach can solve large-size problems. We refer to our proposed estimator as *Convex Entropic Nonparametric* or CENP, because convex programming and entropy are used for purposes of nonparametric estimation.

We restrict our empirical applications to cross-section data. Because time-series data are typically nonstationary, they raise special econometric problems of their own (Beenstock, 1997), which we wish to avoid here. We also avoid other important issues, such as the identification problem first raised by Marschak and Andrews (1944). Finally, we assume that the data on factor inputs are measured without error, and that all the error is in the dependent variable. Had this not been the case it would not have been possible to turn the matter into a convex programming program, which is relatively easy to solve from a nonconvex programming problem, which is very difficult to solve. Therefore we side-step the important issues of stochastic regressors and errors-in-variables. Our main concern is therefore to propose CENP as an econometric methodology and to illustrate its feasibility and performance in the context of production data.

The remainder of the paper is organized as follows. Section 2 describes the maximum likelihood (ML) approach to estimation of the production function. Section 3 shows that the ML estimation problem may be equivalently formulated as a convex program, and includes a brief discussion of numerical procedures used to solve such problems. Section 4 reports numerical results obtained with the data used by Zellner and Ryu (1998). In particular, we show how to evaluate various economic parameters once the CENP was solved. We use these data because we wish to compare CENP with results obtained by Zellner and Ryu's flexible estimators. Specifically, we re-estimate Zellner and Ryu's (1998) NRVES model,⁵ and compare its results with models estimated using our suggested methodology, CENP.

These data do not put CENP through its paces. This is because the data used by Zellner and Ryu happen to include only two factors of production and only 25 data points. Since dimensionality is often a problem in obtaining results with nonparametric estimators, it is important to demonstrate that CENP is feasible when the number of data points is large and when there are more than two factors of production. Therefore, a larger sample is required to assess whether CENP can be used for typical sample sizes. Also, it is desirable to have more than two inputs, if we wish to assess the ability of CENP to cope with the computational complexity of higher dimensionality. To these ends we applied our methodology to US manufacturing data, which contain more than 400 data points and several factors of production, and found that CENP performs well. Section 5 contains concluding remarks. In Appendix A, we prove that the obtained ML estimator is consistent in the sense that it converges with probability one to the true function as the sample size increases to infinity. In Appendix B, we describe our approach to estimating the marginal rates of substitution and the elasticity of substitution used in our numerical analyses.⁶

2. MAXIMUM LIKELIHOOD ESTIMATION

In this section we discuss statistical modeling of the considered data. By \mathbb{R}_+^n (\mathbb{R}_{++}^n) we denote the non-negative (positive) orthant of the n -dimensional vector space \mathbb{R}^n . Let x_1, \dots, x_N be input

⁵ This is model 42 reported in their Table IV, which has the best goodness-of-fit.

⁶ In the nonparametric setting the level sets are piecewise-linear, hence not differentiable, and so calculation of such parameters is not a trivial exercise.

vectors and let y_1, \dots, y_N be the corresponding observed outputs. We assume that the input vectors x_i lie in a convex compact set $\Xi \subset \mathbb{R}_+^n \setminus \{0\}$ and that the outputs y_i are positive, and consider the multiplicative model

$$y_i = \eta_i f(x_i), \quad i = 1, \dots, N \quad (1)$$

Here η_i are positive-valued random variables representing the errors (noise) of the model and $f : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ is viewed as the ‘true’ production function. We assume that $f(\cdot)$ satisfies the following properties: (i) f is concave, (ii) f is component-wise nondecreasing, (iii) $f(0) = 0$, and (iv) $f(x) > 0$ for any $x \in \Xi$. We denote the class of such functions by \mathcal{F} . Note that (1) is equivalent to

$$\ln y_i = \ln f(x_i) + \varepsilon_i, \quad i = 1, \dots, N \quad (2)$$

where⁷ $\varepsilon_i := \ln \eta_i$. We make the following assumptions about the distribution of ε_i :

- (A.1) The random variables ε_i are independently identically distributed (iid) with common probability density function (pdf) $g(\cdot)$.
- (A.2) The pdf $g(\cdot)$ is even (i.e., $g(z) = g(-z)$ for all $z \in \mathbb{R}$), and monotonically decreasing on $[0, +\infty)$ function.

The input variables x_i vary by firm, and, as mentioned in Section 1, are assumed to be measured without error and are assumed to be independent of the errors η_i (and hence ε_i). The multiplicative specification of equation (1) ensures that the outputs y_i are positive as appropriate. The assumption that the errors ε_i are iid (assumption (A.1)) is equivalent, of course, to the assumption that the multiplicative errors η_i are iid, and is quite standard. Assumption (A.2) is rather mild and is made for technical convenience.

Since $g(\cdot)$ is even it follows that the mean of ε_i is zero (provided, of course, that it is finite), and that $g(\cdot)$ is monotonically increasing on $(-\infty, 0]$. It is straightforward then to show that the pdf $p(y)$ of the response variables y_i , conditional on x_i , can be written as follows:

$$p(y) = y^{-1} g(\ln y - \ln f(x_i)) = y^{-1} g\left(\ln\left(\frac{y}{f(x_i)}\right)\right), \quad y > 0$$

and $p(y) = 0$ for $y \leq 0$. Consequently, conditional on $x_i, i = 1, \dots, N$, the likelihood function, of the parameter function ϕ varying over the space \mathcal{F} , can be written as follows:

$$L(\phi) = \prod_{i=1}^N \left[y_i^{-1} g\left(\ln\left(\frac{y_i}{\phi(x_i)}\right)\right) \right] \quad (3)$$

The ML estimate of f is obtained by maximizing $L(\phi)$, or equivalently by minimizing $-\ln L(\phi)$, over $\phi \in \mathcal{F}$. That is, the ML estimate of f is given by an optimal solution of the problem

$$\text{Min}_{\phi \in \mathcal{F}} \sum_{i=1}^N \theta\left(\frac{y_i}{\phi(x_i)}\right) \quad (4)$$

⁷ The notation ‘:=’ means ‘equal by definition’.

where⁸ $\theta(t) \propto -\ln g(\ln t)$, $t > 0$.

The assumed properties on $g(\cdot)$ imply the following properties of the function $\theta(\cdot)$:

- (a) The function $\theta(\cdot)$ is monotonically decreasing on $(0, 1]$ and monotonically increasing on $[1, +\infty)$, and hence has its minimum at $t = 1$.
- (b) $\theta(t) \rightarrow +\infty$ as $t \rightarrow 0$ or $t \rightarrow +\infty$.
- (c) $\theta(t^{-1}) = \theta(t)$ for any $t > 0$.

Property (a) follows directly from assumption (A.2). Property (b) holds since $g(z) \rightarrow 0$ as z tends to $+\infty$ or $-\infty$. Property (c) follows from the definition of $\theta(\cdot)$ and since $g(\cdot)$ is an even function. We may consider the following examples of the function $\theta(\cdot)$. If ε_i have a normal distribution (with zero mean), then $\theta(t) = (\ln t)^2$, and if $g(z) \propto e^{-(e^z + e^{-z})}$, then $\theta(t) = t + t^{-1}$.

Since the function $\theta(\cdot)$ satisfies the above conditions (a), (b) and (c), it attains its minimum at the point $t = 1$. Therefore, if in addition $\theta(\cdot)$ is smooth, then $\theta'(1) = 0$ and $\theta''(1) \geq 0$. Consequently $\theta(\cdot)$ has the following second-order Taylor expansion at $t = 1$:

$$\theta(t) = a + b(t - 1)^2 + o((t - 1)^2) \quad (5)$$

where $a = \theta(1)$ and $b = \frac{1}{2}\theta''(1) \geq 0$. Therefore, for errors $\eta_i = y_i/f(x_i)$ close to one, estimation procedures based on solving (4) for two different functions θ are *asymptotically equivalent*, as long as $\theta''(1) > 0$ for both of them. In particular, for $\theta(t) := (\ln t)^2$ and $\theta(t) := t + t^{-1}$ the second derivative at $t = 1$ is 2, and hence the corresponding estimation procedures are asymptotically equivalent. Note, however, that from a computational perspective the function $\theta(t) = t + t^{-1}$ is preferred since it is *convex*.

Let us observe at this point that typically the optimization problem (4) has many (infinitely many) optimal solutions. It is possible to show, however, that under mild regularity conditions any optimal solution \hat{f}_N of (4) converges w.p.1 to the true function f as $N \rightarrow \infty$. We will discuss this consistency property of ML estimators in the Appendix.

3. NONPARAMETRIC ESTIMATION AS A CONVEX PROGRAM

In this section we discuss an approach to a numerical solution of the ML optimization problem. We start by giving a characterization of functions from the class \mathcal{F} which take given values at the input points.

Definition 3.1 A data set $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^N$ of input–output pairs is said to be *concave-representable* if there exists a function $\Phi \in \mathcal{F}$ for which $\Phi(x_i) = y_i$.

It is natural to view the function Φ in the above definition as defined on the set

$$\mathcal{X} := \text{conv}\{x_1, \dots, x_N\} + \mathbb{R}_+^n \quad (6)$$

⁸ The notation ' $\theta(t) \propto$ ' means that $\theta(\cdot)$ is proportional (i.e., is equal up to a positive multiplicative constant) to a given function.

where $\text{conv}\{x_1, \dots, x_N\}$ denotes the convex hull of the input vectors. That is, $x \in \chi$ if there exist $\lambda_i \geq 0, i = 1, \dots, N$, such that $\sum_{i=1}^N \lambda_i = 1$ and $x \geq \sum_{i=1}^N \lambda_i x_i$.

Given a dataset \mathcal{D} and a vector $\sigma \in \mathbb{R}_+^N$, let $\mathcal{D}(\sigma)$ denote the dataset given by $\{(x_i, \sigma_i y_i)\}_{i=1}^N$, and let $\Sigma(\mathcal{D})$ denote the set of all such vectors σ for which $\mathcal{D}(\sigma)$ is concave-representable. Obviously, \mathcal{D} is concave-representable if and only if the vector $(1, 1, \dots, 1) \in \Sigma(\mathcal{D})$. Using property (c) of function $\theta(\cdot)$, problem (4) can be written in the form

$$\text{Min}_{\sigma} \sum_{i=1}^N \theta(\sigma_i) \text{ subject to } \sigma \in \Sigma(\mathcal{D}) \tag{7}$$

We now turn to characterizing the implicit constraint ‘ $\sigma \in \Sigma(\mathcal{D})$ ’ into a set of explicit constraints so that the problem (7) can be computationally solved.

3.1. Characterization of Concave-Representable Datasets

For a given dataset \mathcal{D} and $x \in \chi$, denote by $\Phi^*(x)$ the optimal value of the following linear programming problem:

$$\Phi^*(x) := \text{Max}_{\lambda} \sum_{i=1}^N \lambda_i y_i \text{ subject to } \sum_{i=1}^N \lambda_i x_i \leq x, \sum_{i=1}^N \lambda_i = 1, \lambda_i \geq 0, i = 1, \dots, N \tag{8}$$

Lemma 3.1 *If \mathcal{D} is concave-representable, then Φ^* is its minimal representation, namely, $\Phi(\cdot) \geq \Phi^*(\cdot)$ on χ for any other representation Φ .*

Proof For every $x \in \chi$, the feasible set of problem (8) is nonempty. Clearly, Φ^* is non-negative, nondecreasing and finite-valued. A straightforward argument shows that Φ^* is also concave. Hence, $\Phi^* \in \mathcal{F}$. Let Φ denote a representation for \mathcal{D} . Pick an $x \in \chi$ and a feasible vector λ in (8). Since Φ is both nondecreasing and concave:

$$\Phi(x) \geq \Phi\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i \Phi(x_i) = \sum_i \lambda_i y_i \tag{9}$$

Since (3.4) holds for all feasible λ the minimality of Φ^* immediately follows. □

Proposition 3.1 *The dataset \mathcal{D} is concave-representable if and only if $\Phi^*(x_k) = y_k$ for all $k = 1, \dots, N$.*

Proof The if part follows immediately from the concavity of Φ^* . Since $\lambda_i = 0, i \neq k$, and $\lambda_k = 1$ is feasible for (8) we have that $\Phi^*(x_k) \geq y_k$ for each k . The converse now immediately follows from Lemma 3.1, since $y_k = \Phi(x_k) \geq \Phi^*(x_k) \geq y_k$. □

Remark The characterization given by Proposition 3.1 is essentially the same as presented in Banker and Maindiratta (1992). Our proof establishes the minimality of Φ^* to obtain a concise proof.

3.2. Dual Representation

Let $\Phi_\sigma^*(x)$ denote the function defined as the optimal value of problem (8) in which each y_k has been replaced with $\sigma_k y_k$. Theorem 3.1 demonstrates that $\sigma \in \Sigma(\mathcal{D})$ if and only if $\Phi_\sigma^*(x_k) = \sigma_k y_k$ for all k . However, as it stands, this equivalence is not directly helpful for the purpose of solving (7), since σ appears on both sides of the identity. Fortunately, the optimization problem that defines $\Phi^*(x_k)$ is a linear program, and so we may appeal to the duality theory of linear programming. For $x = x_k$ the dual linear program to (8) is

$$\text{Min}_{p \geq 0} p^T x_k + \pi \quad \text{subject to } p^T x_i + \pi \geq y_i, i = 1, \dots, N \quad (10)$$

When \mathcal{D} is concave-representable, (10) has an optimal solution (p_k, π_k) , which satisfies the following equations:

$$p_k^T x_i + \pi_k \geq y_i, i = 1, \dots, N, \quad (11)$$

$$p_k^T x_k + \pi_k = y_k \quad (12)$$

Note that if (11) and (12) hold, then the optimal value for the dual linear program is obviously y_k , which must equal $\Phi^*(x_k)$ by linear programming duality. Thus, we have established the following result.

Corollary 3.1 *The set \mathcal{D} is concave-representable if and only if for each $k = 1, \dots, N$, there exist $p_k \geq 0$ and π_k that satisfy (11) and (12).*

As a direct consequence of Corollary 3.1, problem (7), and hence problem (4), can be formulated as the following optimization problem:

$$\begin{aligned} & \text{Min}_{\sigma \geq 0, p \geq 0, \pi} \sum_{k=1}^N \theta(\sigma_k) \\ & \text{subject to } p_k^T x_i + \pi_k \geq \sigma_i y_i, i, k = 1, \dots, N, \\ & \quad p_k^T x_k + \pi_k = \sigma_k y_k, k = 1, \dots, N \end{aligned} \quad (13)$$

The above problem has convex objective function and linear constraints, and hence is a convex programming problem.

We note that elimination of π_k in equations (11) and (12) shows that \mathcal{D} is also concave-representable if and only if for each k there exist $p_k \geq 0$ for which

$$y_k - p_k^T x_k \geq y_i - p_k^T x_i, \quad i = 1, \dots, N \quad (14)$$

Equation (14) shows that each p_k defines a supergradient of the concave function Φ^* at (x_k, y_k) . The use of supergradients provides an alternative, direct means to establish the existence of (11) and (12). In particular, given a set of supergradients that satisfy (14) one may define

$$\phi(x) := \min_{1 \leq k \leq N} \{y_k + p_k^T (x - x_k)\}$$

The function $\phi(\cdot)$, being the minimum of a finite collection of linear functions, is concave, and it is not difficult to show it is a valid representation of the data—see Matzkin (1999, Lemma 1) for details. However, this representation is dependent on the particular choice of supergradients, and is therefore *not* unique. The above duality approach taken here shows that the optimal value of (10) gives the corresponding *minimal* representation. We use this minimal representation to estimate the returns to scale and the elasticity of substitution.

Equation (14) has a natural economic interpretation very much in the spirit of the *Revealed Preference* literature. Normalizing the price on output to be 1 these equations simply state that \mathcal{D} is concave-representable if and only if for each ‘firm’ k there exist prices on inputs for which the observed input–output choice (x_k, y_k) maximizes the firm’s profit. Similar types of equations will exist to characterize the technology depending on what one assumes about what data (inputs, outputs, prices, cost, profits, etc.) are observed (see Varian, 1982, 1983, 1984).

3.3. Entropic Distance

The objective function of problem (13):

$$\Theta(\sigma|1) := \sum_{i=1}^N \theta(\sigma_i) \quad (15)$$

may be viewed as a *distance* between a given vector of adjustments $\{\sigma_1, \sigma_2, \dots, \sigma_N\}$ and the *ideal* vector of adjustments given by $1 = \{1, 1, \dots, 1\}$. The properties on θ imply that Θ fits the notion of *entropic distance* introduced first by Csiszar (1967). Ostensibly, other entropic distance functions could be used in (15); for a discussion of such functions, see Ben-Tal *et al.* (1989). Due to its connections with entropic distance and convex programming, we have termed our proposed *ML* estimator as formulated in (13) *convex entropic nonparametric* (CENP). Note that the functions $\phi_n(\sigma) = \sigma^n + \sigma^{-n} - 2$ for $n = 1, 2, \dots$ are all entropic distance functions. In the present paper we experimented with three different entropic functions:

$$\begin{aligned} \text{(a) } \theta_1(\sigma) &= \phi_1(\sigma) \\ \text{(b) } \theta_2(\sigma) &= \phi_2(\sigma) \\ \text{(c) } \theta_3(\sigma) &= \phi_2(\sigma) - 4\phi_1(\sigma) - 8 \end{aligned} \quad (16)$$

3.4. Convex Programming Algorithms

We used the LMI ToolBox to solve our separable convex programming formulation \mathcal{D} for small-sized problems ($N < 50$). It is a standard ToolBox supplied with the MATLAB software package. We also used GAMS MINOS, which is a commercially available optimization package. While LMI was found to be inefficient for larger problems ($N > 130$), GAMS MINOS solved efficiently, within a few minutes, problems with datasets containing up to 200–230 points. It was not, however, possible to solve a very large problem ($N > 400$) using either software package.

There are a number of specialized algorithms to solve problems like \mathcal{D} that exploit the separability and strict convexity of the objective function (consult, for example, Bazarra *et al.*, 1993). A simple approach, which is conceptually easy to understand and not difficult to formulate

and implement, is based on constructing a finite supporting hyperplane to provide a piecewise linear approximation of θ that bounds it from below.

Problem (13) can be formulated as

$$\begin{aligned} \text{Min}_{\sigma \geq 0, p \geq 0, \pi, \gamma} \quad & \sum_{k=1}^N \gamma_k \\ \text{subject to} \quad & \theta(\sigma_k) \leq \gamma_k, k = 1, \dots, N, \\ & p_k^T x_i + \pi_k \geq \sigma_i y_i, i, k = 1, \dots, N, \\ & p_k^T x_k + \pi_k = \sigma_k y_k, k = 1, \dots, N \end{aligned} \quad (17)$$

Choose a set of ‘grid points’ $\sigma_{k\ell}$, $\ell = 1, \dots, M$, $k = 1, \dots, N$, preferably with values around and closed to 1, and define $\theta_{k\ell} = \theta(\sigma_{k\ell})$. Due to convexity, we can replace the *convex* inequality (3.12) for each k with the set of *linear* inequalities given by

$$\theta_{k\ell} + [d\theta(s)/ds]_{s=\sigma_{k\ell}} \times (\sigma - \sigma_{k\ell}) \leq \gamma_k, \ell = 1, \dots, M$$

The corresponding new problem is a linear program. Since the piecewise linear approximation supports the original objective function, the solution of the linear program, $\sum_{k=1}^N \gamma_k^*$, serves as a lower bound for the true solution of problem (13). If (σ^*, γ^*) is a solution of the linear program, then $\sum_{k=1}^N \theta(\sigma_k^*)$ serves as an upper bound for the true solution of problem (13).

4. RESULTS

In this section we report an empirical application of CENP to data used by Zellner and Ryu (1998), which were collected in 1957 for a cross-section of 25 US states. This dataset consists of two inputs: man-hours (measured in millions) and capital services (measured in millions of US dollars) used to manufacture a single output, the value-added of transportation equipment (measured in millions of US dollars). We use this dataset because we could re-estimate the various models estimated parametrically by Zellner and Ryu, and compare them with models estimated nonparametrically using our suggested methodology, CENP. The data have been normalized by Zellner and Ryu by the number of establishments in each state.⁹

As mentioned in Sections 1 and 2, CENP assumes that all the measurement error is in the y 's. Therefore the x 's are assumed to be measured without error and are not stochastic. If, however, the x 's and the y 's contained measurement error CENP estimates would most probably be biased and inconsistent. CENP also assumes that measurement errors in the y 's are independent. If states happened to experienced common shocks in 1957 the errors would not be independent; they will be positively correlated. For example, if the transportation equipment sector happened to be booming in 1957 total factor productivity would likely be higher in each state. This would not matter if this effect was identical in each state, because y in each state would increase by the same proportion. However, if this effect is heterogeneous positive error dependence would be induced. Error dependence creates inefficiency but not inconsistency in linear regression models,

⁹ The data are available on the website of the *Journal of Applied Econometrics* (www.econ.queensu.ca/jae/).

and it induces both inefficiency and inconsistency in nonlinear regression models. Most probably positive error dependence also induces inconsistency in CENP. Therefore error independence is important for CENP just as it is for parametric estimators, such as those suggested by Zellner and Ryu. These restrictions obviously qualify the results that we are about to report.

In this section, we shall check the sensitivity of CENP to the three chosen entropic distance functions (16), compare the results obtained by parametric methods with those obtained by CENP, and discuss how estimates of production technologies obtained by CENP may be used to calculate a variety of economic phenomena, such as elasticity of substitution and returns to scale.

4.1. Estimation by CENP

To check the sensitivity of the solution to various *entropic* distances we have solved it for each function $\theta_i(\sigma)$, $i = 1, 2, 3$, in (16). The *quality* of each estimation function is measured by the percentage root mean square error $\left(\text{RMSE} = \sqrt{\frac{1}{N} \sum \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \right)$. The three entropic functions generate similar estimated values for value added, and RMSE is 0.1336 for the first two functions and is slightly larger for the third. The 95% confidence interval for RMSE obtained by the bootstrapping procedure described later is 0.130–0.141, in which case CENP estimates RMSE quite precisely despite the small sample size.

A natural question arises whether it is possible to verify the assumption of monotonicity and concavity of the ‘true’ response function. That is, is it possible to test the hypothesis that $f(\cdot)$ is monotone and/or concave assuming that for given data the model (1) holds for some function $f(\cdot)$? Such questions were studied in the literature on testing shape (or curvature) constraints (see Robertson *et al.*, 1988). A parametric approach to testing monotonicity (concavity) can be based on the likelihood ratio method and the so-called chi-bar-squared distributions (see Doherty *et al.*, 2002). An interesting nonparametric approach and a survey of relevant literature can be found in Abrevaya and Jiang (2002). All these tests are asymptotic and cannot be reasonably applied to the small datasets analyzed here.

In the meantime we consider whether the null hypothesis of a monotonic and concave production technology is rejected by the data. The probability of rejection naturally increases with RMSE. If $\hat{\sigma}_i = 1$ for all the data points then the data satisfy the null hypothesis precisely and $\text{RMSE} = 0$. Strictly speaking, any $\hat{\sigma}_i \neq 1$ would constitute a rejection of the null hypothesis. However, if the data happen to contain measurement error such rejections of the null may not be statistically significant. We follow Varian (1985) in asking how large measurement error in y would have to be for RMSE not to reject the null. Let ν denote the unknown variance of measurement error, let $\text{ESS} = N(\text{RMSE})^2$ denote the error sum of squares generated by CENP, and define $S = \frac{\text{ESS}}{\nu}$. If the errors generated by CENP happen to be normally distributed, Varian suggests rejecting the null if $S > \chi_{p,N}^2$, where $\chi_{p,N}^2$ denotes the critical value of χ^2 at probability p with N degrees of freedom. S is larger the greater is ESS relative to the unknown variance of measurement error.

Since ν is unknown, Varian suggests calculating $\bar{\nu} = \frac{\text{ESS}}{\chi_{p,N}^2}$ as the upper-bound for ν below which we would reject the null. The smaller is $\bar{\nu}$ the more reasonable it would be not to reject the null. For example, in the case of $\theta_1(\sigma)$ the $\text{ESS} = 0.4462$ and $\chi_{0.05,25}^2 = 14.61$, in which case $\bar{\nu} = 0.0305$. If the true measurement error variance is less than 0.0305, or 3.05%, we should reject the null hypothesis, but if it is larger than 3.05% we should accept the null. Since 3.05% is a

modest error variance we are inclined not to reject the null of concavity. Indeed, when two outliers are omitted (Kentucky and New York) \bar{v} falls substantially.

Varian's test would only be valid if the CENP residuals happened to be normally distributed and independent. In Varian (1985) the data are time series, in which case serial correlation would invalidate the assumption of independence. Varian did not check whether the model errors were independent or normally distributed. Because we use cross-section data it is more reasonable to assume that the errors are independent although, as already noted, common shocks may induce positive error dependence in cross-sections. We use the Jarque–Bera statistic (JB), which has a chi-square distribution with 2 degrees of freedom, and which does not require the errors to be independent, to check whether the CENP residuals happen to be normally distributed.

The mean error generated by case $\theta_1(\sigma)$ is 0.0099, or almost 1%, which is not significantly different from zero. Note that although we did not impose the restriction of a zero mean error, CENP generates this result spontaneously. The JB statistic is 2.52, which is less than the critical value of $\chi_{0.05,2}^2 = 5.99$. Therefore, the CENP residuals seem to be normally distributed, which suggests that in the present case Varian's test is appropriate. When the two outliers mentioned above are omitted the case for normality is even stronger. Note that CENP does not assume normality; it is a spontaneous result.

There are a number of problems with Varian's test. The first is that it assumes that the population errors are normally distributed. It seems rather odd to propose a parametric testing procedure in a nonparametric context. The second is that v is unknown so that the test is inevitably subjective. Third, the number of degrees of freedom is less than N if $\hat{\sigma}_i$ are not independent. The measurement of degrees of freedom is not straightforward in nonparametric estimation, and it remains a problem here. However, we suggest the Kolmogorov–Smirnov test (KS) as a nonparametric alternative to Varian's test. The advantage of KS is that it makes no parametric assumptions about the distribution of the population errors. A disadvantage is that KS assumes that measurement error in the y 's is independent. Gleser and Moore (1983) discuss the implications of positive error dependence for KS. Not surprisingly they show that positive dependence adversely affects the level of significance and power because positive dependence is confounded with lack of fit.

The KS test statistic is $D(N) = \max |G(\hat{\sigma}_i - 1) - F(\hat{\sigma}_i - 1)|$, where $F(\cdot)$ is the observed cumulative distribution of the estimated model errors and $G(\cdot)$ is the hypothesized cumulative distribution. The critical values of $D(25)$ range between 0.21 at $p = 0.2$ to 0.32 at $p = 0.01$. If, for example, we assume that $G(\cdot)$ is cumulative normal with variance equal to 0.01 (1%) the calculated value for D is 0.121, which falls well short of its critical value. We therefore cannot reject the null hypothesis that $\sigma_i = 1$. Hypothesizing a lower variance naturally increases the calculated value for D and raises the chances of rejecting the null. For example, if the variance is only $\frac{1}{2}\%$ instead of 1% $D = 0.173$, which still falls short of its critical value. Most probably these results are sufficiently strong despite the adverse effect of positive dependence on their statistical power.

4.2. CENP vs. Parametric Methods

To compare estimates obtained by CENP to parametric estimates requires replication of the results reported by Zellner and Ryu (1998). Their most flexible parametric model (NRVES), which also happens to have the best goodness-of-fit, uses a Box–Cox transformation that allows the

elasticity of substitution to vary both with respect to factor proportions and with respect to scale.¹⁰ We use NRVES to represent the generalized production function approach parametrically. For completeness, we also estimate a translog model, despite the fact that it does not necessarily possess neoclassical properties, and for which reason Zellner and Ryu did not estimate it. We continue to use percentage root mean square to measure *goodness-of-fit*.

For comparison purposes, we calculate the outputs of CENP derived from the three entropy functions. The results are reported in Table I. We wish to stress that the comparisons are not intended as a horse race in which the winner takes all. Nevertheless, it is reassuring to note that CENP performs well against flexible parametric alternatives. This is to be expected because parametric estimators are more parsimonious than their nonparametric counterparts. Unfortunately there is no agreed way to correct nonparametric estimates for degrees of freedom since the concept of degrees of freedom is foreign to nonparametric statistics. Nevertheless, Hastie and Tibshirani (1990, Ch. 3) suggest a heuristic measure of ‘degrees of freedom equivalence’ for smoothers, which varies inversely with the degree of smoothing. Although we recognize the importance of the issue we have not been able to calculate formal d.f. equivalences for CENP. However, informally we do not think that CENP is expensive in terms of degrees of freedom since, like NRVES, it imposes only two restrictions upon the data: monotonicity and concavity. And like NRVES, CENP does not restrict returns-to-scale and elasticity of substitution to be constant, so the number of degrees of freedom used by CENP is most probably similar to that of NRVES. Therefore the comparison between CENP and NRVES in Table I is not invidious.

To test for robustness, we sequentially omitted one data point and estimated the production function based only on 24 observations. We then estimated the value of the production function at the deleted data point. The results of this *cross-validation* test are provided in Table I. RMSE naturally increases. For CENP it increases from 0.135 to 0.213 and the advantage of CENP over NRVES and translog¹¹ narrows but does not disappear.

4.3. Applications of CENP

In this section we apply the methods described in Appendix B to calculate such parameters of interest as ES and RTS from the CENP model estimates reported in Section 4.1. We begin by plotting the isoquants or level sets associated with several states (see Figure 1). By construction these sets are convex and piecewise linear, except for those corresponding to the smallest and largest adjusted outputs, Florida and Michigan, whose isoquants collapse onto a single point. The isoquants depicted in Figure 1 have a vertical segment on the left-hand side and a horizontal

Table I. Cross validation tests

	CENP	Translog	NRVES
RMSE—estimation	0.135	0.1514	0.1537
RMSE—cross-validation	0.2128	0.2137	0.2142

¹⁰ Zellner and Ryu do not name a ‘preferred’ model out of the numerous models they estimated. However, they name several models as inappropriate. We choose NRVES not merely on the grounds of goodness-of-fit, but also because Z&R mention that it has desirable properties.

¹¹ Of course, the translog model may be inconsistent with neoclassical production theory, and therefore undesirable even had its goodness-of-fit been superior.

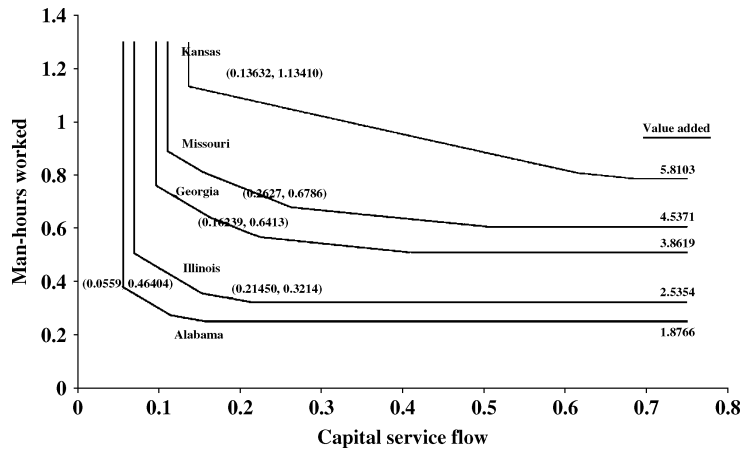


Figure 1. Level sets of five states

segment on the right-hand-side, which reflect the support of the data. This means that the data do not permit CENP to extrapolate beyond the observations. Data points in the center of the data set naturally have more linear segments because we learn more about the production technology for such data points, and less at the extremes of the data. The largest number of segments is five (excluding the vertical and horizontal segments). Within any internal segment we cannot say how the production technology varies, because there are not sufficient data to guide us. Had there been more data points the derived isoquants would have had more segments, and would have appeared more continuous. In the limit the isoquants would tend to be continuous.

The isoquants plotted in Figure 1 are estimates, which are subject to estimation error. It is therefore natural to ask about the confidence intervals for the estimates that we report. Some non-parametric estimators have analytical expressions for confidence intervals (e.g. Härdle, 1990) for the kernel estimator. Since there is no analytical expression for calculating confidence intervals for CENP we bootstrap them for various parameters of interest. We found 100 bootstraps to be sufficient for our purposes in the sense that confidence intervals tended to converge on some constant value. This number is quite low compared to what might have been expected from Andrews and Buchinsky (2001). The 95% confidence intervals for value added in Figure 5 are 5.7281–5.9443 for Kansas, 4.6369–4.7431 for Missouri, 3.8911–3.9899 for Georgia, 2.5579–2.5963 for Illinois, and 1.8501–1.8707 for Alabama. The upper confidence intervals vary from 0.55% of the mean in the case Alabama to 1.85% in the case of Kansas. These confidence bands are quite small and seem to be scale dependent. These calculations show that CENP estimates value added to a reasonable degree of precision.

Next we examine returns-to-scale. In Figure 2, we plot the returns-to-scale generated by the CENP model at three different levels of labor intensity. The point marked 'Alabama' is the coordinate for capital services and output in Alabama where the capital–labor ratio is 0.12. Imagine a ray from the origin that passes through the indicated coordinates on the isoquant for Alabama in Figure 1. This ray would intersect higher and lower isoquants (not shown). Along the Alabama schedule in Figure 2 the capital–labor ratio is held constant at 0.12. Since the vertical axis measures the logarithm of output and the horizontal axis measures the logarithm of capital services, the returns-to-scale at any point is the derivative of the (plotted) function at this point.

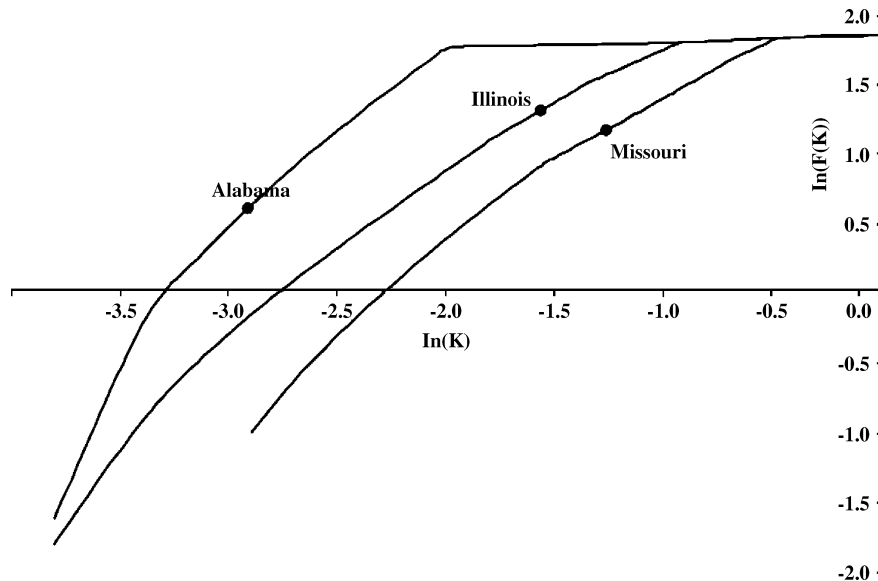


Figure 2. Returns to scale and labor intensity

Returns-to-scale are increasing if the derivative is greater than one and decreasing if the derivative is less than one. As each of the returns-to-scale functions are *almost* piecewise-linear, the slopes of each line segment can easily be calculated.

At the point marked 'Alabama' the derivative is greater than unity, in which case returns-to-scale are increasing at this point; i.e., the slope of the line exceeds 45 degrees. The 95% confidence interval obtained by bootstrapping is 1.302–1.799; therefore we can rule out the possibility of constant returns-to-scale, although returns-to-scale are not estimated precisely. The slope of the Alabama schedule in Figure 2 becomes steeper as the scale of output is reduced. This implies that when the capital–labor ratio is 0.12 returns-to-scale are greater the lower the level of output. Indeed, while returns-to-scale decrease at higher output levels, the Alabama schedule shows that at the end of the support of the data there are still increasing returns-to-scale.

Along the Illinois schedule in Figure 2 the capital–labor ratio is held constant at 0.67, which is the ratio for Illinois. Production in Illinois was considerably more capital intensive than in Alabama. At the point marked 'Illinois' the derivative of the schedule is less than unity; hence there are diminishing returns-to-scale at this point. The 95% confidence interval obtained by bootstrapping is 0.950–0.968; therefore returns-to-scale are clearly diminishing, and in this case RTS is estimated precisely. The Illinois schedule shows, however, that at lower levels of output returns-to-scale are increasing. Hence, at some point returns-to-scale start decreasing, and they continue to decrease with the scale of output. Finally, the Missouri schedule refers to a capital–labor ratio of 0.39, as in Missouri. At the point marked 'Missouri' there are decreasing returns-to-scale. The bootstrapped confidence interval is 0.885–0.899, so we can be sure that while returns-to-scale decrease in Missouri, they decrease more strongly than in Illinois. Note also that RTS is estimated quite precisely. However, at lower scales of production there are increasing returns-to-scale in Missouri.

We turn next to the elasticity of substitution. We calculate the elasticity of substitution between labor and capital at different levels of output and at different levels of labor intensity. Figure 3 plots

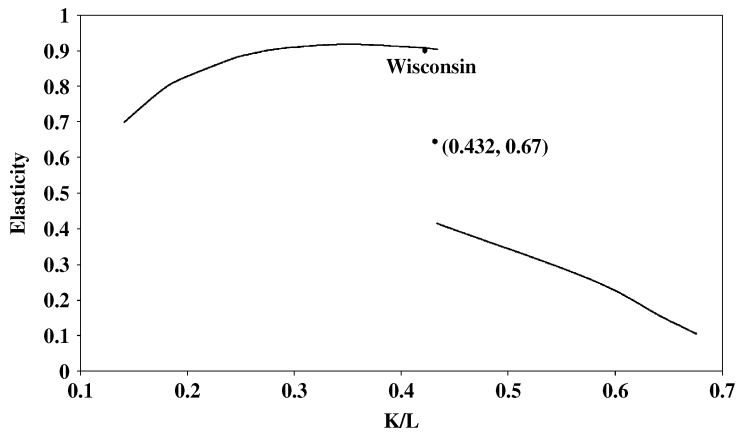


Figure 3. The Elasticity of substitution along the isoquant

the relationship between the elasticity of substitution and the labor–capital ratio that is generated along the level sets corresponding to the level of output in Wisconsin. Because the calculation of the elasticity of substitution involves the second derivative of the production function, and because the isoquants are piecewise linear, the elasticity of substitution is not always defined—hence the discontinuities observed in Figure 3. There are smaller discontinuities too in Figure 3, but these are too small to be observed by the naked eye.

The elasticity of substitution is less than unity, but increases with the labor–capital ratio. The 95% confidence interval obtained by bootstrapping for the elasticity of substitution is 0.853–0.913 in Wisconsin and 0.384–0.428 in Kentucky. Therefore we can be sure that the elasticity of substitution is less than unity. Indeed, CENP estimates this parameter quite precisely. The estimated relationship between the capital–labor ratio and the elasticity of substitution is not monotonic, and the elasticity of substitution eventually decreases.

In a further exercise we use the CENP model to calculate the elasticity of substitution for each of the 25 states as a function of their observed capital–labor ratios. These capital–labor ratios range between 0.2 and 0.7 and the elasticities of substitution range between 0.2 and 1.2. Figure 4 shows that the elasticity of substitution varies quite substantially at given capital–labor ratios. The reason for this is that the elasticity of substitution depends on the scale of output as well as factor proportions. If one ignores the three observations at the bottom left of Figure 4, it suggests that the elasticity of substitution tends, on the whole, to vary inversely with the capital–labor ratio.

Finally, Figure 5 plots the distribution of the estimated elasticity of substitution obtained by bootstrapping. At each bootstrap we calculate for each state the elasticity of substitution at the means of the data for labor and capital. The estimated elasticity of substitution turns out to have an asymmetric distribution. The mode of the distribution is 0.7, but with probability 0.15 the elasticity of substitution is only 0.2. The probability that the elasticity of substitution exceeds unity is only about 0.15. Figure 5 creates the misleading impression that CENP does not estimate ES very precisely. However, the confidence intervals reported for Wisconsin and Kentucky show that CENP estimates ES quite precisely.

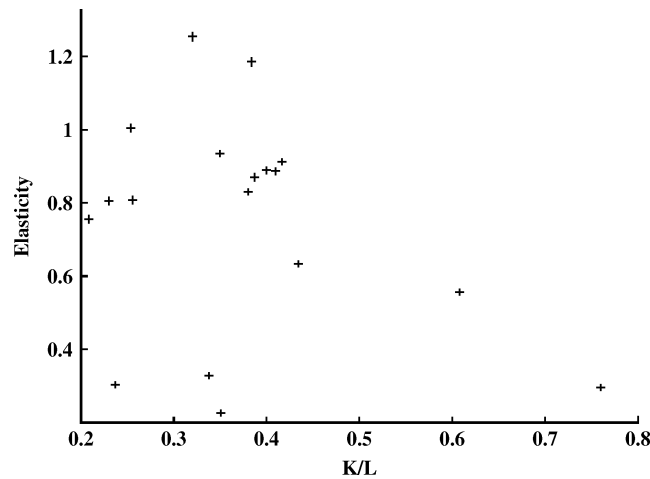


Figure 4. The elasticity of substitution evaluated at each state

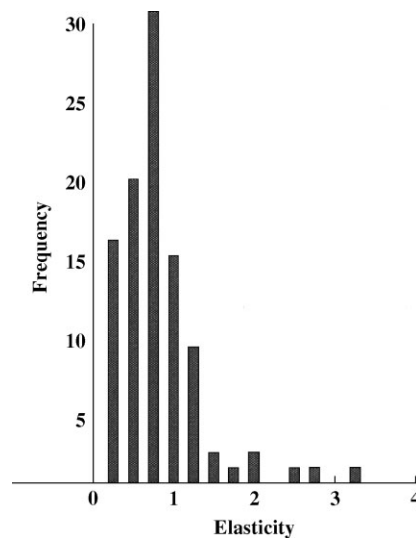


Figure 5. The bootstrapped distribution of the elasticity of substitution

4.4. CENP with a Large Dataset

The empirical application of CENP reported in Section 4.1 had only 25 data points and 2 covariates. As the number of data points and covariates increases the dimensionality of the estimation problem naturally grows. It is therefore reasonable to ask how the ‘curse of dimensionality’, which as discussed, for example, by Härdle (1990) arises in other forms of nonparametric estimation, affects CENP.

In Section 3.4 we suggested a linear programming procedure for approximating the original objective function. We applied this procedure to a much larger problem in which the number of

data points is 448 and the number of covariates is 3. In this problem the dependent variable is value added in 448 manufacturing sectors in the USA (downloaded from the website of NBER) in 1954, and the covariates consist of capital, labor and raw materials as factor inputs.

We have solved the linear program with $M = 5$ so the number of grid points was $448 \times M = 2240$. This problem was solved using the CPLEX software for linear programming. It took less than 30 minutes to obtain a near-optimal solution with a provable error bound of less than 10%. Of course, solution quality increases with larger M . It is possible to sequentially improve quality by simply adding the supporting hyperplane to each point $(\sigma_k, \theta_k(\sigma_k))$ after each iteration, and then stopping when the ratio of the upper to lower bounds is sufficiently close to 1. In our experimentation a fixed value of $M = 5$ proved adequate. Computation time for CENP depends merely on the number of data points through the number of the constraints. If the number of data points increases from N to $(N + 1)$, the number of constraints increases by $(2N + 1)$. However, computation time is virtually insensitive to the number of inputs. If the number of inputs increases by 1 the number of variables increases by 1, while keeping the constraints unchanged.

5. CONCLUSIONS

More than 30 years ago Hanoch and Rothschild (1972) suggested a nonparametric methodology for testing the predictions of production theory. They saw that their methodology could, in principle, be used to provide nonparametric estimates of the production function. However, they did not envisage that this was practically feasible. Rather they saw their approach in the less ambitious role as a screening device, as a technique for inspecting the data for coherence with production theory. During the early 1980s Varian returned to this theme, but stopped short of proposing a nonparametric estimator of production and other functions in economics. Instead, he limited himself to screening the data along the lines suggested by Hanoch and Rothschild.

The last two decades have been virtually silent on the issue. Exceptions include Matzkin (1991, 1993, 1994, 1999) and Banker and Maindiratta (1992), who suggested a way to estimate production functions nonparametrically. Banker and Maindiratta did not, however, show that their proposal is feasible, and we doubt that it is for reasons stated in Section 1. Matzkin (1999) proposed a specialized algorithm that shows promise for small-size problems. In any case, we are unaware of subsequent applications, successful or otherwise. Our main contribution has been to show how the Hanoch–Rothschild methodology can be turned into a nonparametric estimator from a mere screening device. Our solution to the problem is based on convex programming and entropic methods. Hence, we refer to our estimator as CENP. Our most important contribution therefore is a practical solution to an old problem. CENP joins the small but expanding literature on nonparametric estimation subject to shape constraints.

To demonstrate the feasibility of CENP we reported two empirical applications. In the first the dataset was small, containing only 25 data points. We chose this dataset so that we could compare results obtained by CENP with flexible functional forms estimated parametrically by Zellner and Rhu (1998). CENP outperforms its most flexible parametric rivals because such parameters as returns-to-scale and elasticity of substitution vary empirically in a way that even very flexible functional forms find difficult to accommodate. In the second empirical application there were more than 400 data points. Our intention was to demonstrate the feasibility of CENP when the dataset is relatively large. We show that this is indeed the case.

Since analytical estimates of parameter uncertainty are not available for CENP, we bootstrapped CENP in order to estimate confidence intervals for key parameters. Even though the sample size was small, we found that CENP estimates value added and parameters such as returns-to-scale and elasticity of substitution quite precisely. Therefore, CENP is not only feasible as a nonparametric estimator subject to shape constraints, but it also lends itself to testing hypotheses regarding such matters as the constancy of returns-to-scale and the size and variability of the elasticity of substitution.

CENP is computer intensive. However, the falling cost of computing does not explain the timing of the present research. On the other hand, the falling cost of computing makes CENP more attractive than it might have been a decade ago. We see CENP as part of the econometrician's toolkit, which has applications in other fields of economic inquiry apart from production where monotonicity and concavity are relevant shape constraints.

APPENDIX A

In this appendix we show that under mild regularity conditions the maximum likelihood optimization procedure produces a consistent estimator of the true function $f(x)$. In addition to the assumptions (A.1) and (A.2), specified in Section 2, we assume that the distribution of the random variables ε_i is log-concave. That is:

(A.3) The function $h(z) := -\ln g(z)$ is strictly convex on \mathbb{R} .

Note that $h(z) := \theta(e^z)$ and $\theta(t) = h(\ln t)$, and that $h(\cdot)$ is an even function since $g(\cdot)$ is even. For example, the functions $\theta(t) := (\ln t)^2$ and $\theta(t) := t + t^{-1}$ satisfy the above assumption (A.3). We also assume that all involved expectations do exist.

Lemma A.1 Under the assumptions (A.1)–(A.3), the function $\psi(t) := \mathbb{E}\{\theta(t\eta)\}$, where $\ln \eta \sim g(\cdot)$, attains its minimum, over \mathbb{R}_{++} , at the point $t = 1$, and this minimizer is unique.

Proof We have that

$$\theta(t\eta) = h(\ln(t\eta)) = h(\ln \eta + \ln t)$$

and hence $\psi(t) = \mathbb{E}[h(\varepsilon + \tau)]$, where $\varepsilon \sim g(\cdot)$ and $\tau := \ln t$. Since functions $h(\cdot)$ and $g(\cdot)$ are even, we have

$$\begin{aligned} \mathbb{E}[h(\varepsilon + \tau) - h(\varepsilon)] &= \int_{-\infty}^{+\infty} [h(z + \tau) - h(z)]g(z)dz \\ &= \int_0^{+\infty} [h(z + \tau) + h(z - \tau) - 2h(z)]g(z)dz \end{aligned}$$

Moreover, since $h(\cdot)$ is strictly convex, we have $h(z + \tau) + h(z - \tau) - 2h(z) > 0$ for all z and $\tau \neq 0$. It follows that $\mathbb{E}[h(\varepsilon + \tau)] > \mathbb{E}[h(\varepsilon)]$ for any $\tau \neq 0$, and hence $\psi(t)$ attains its minimum when $\ln t = 0$, i.e., $t = 1$, and this minimizer is unique. \square

Recall that it was assumed that $x_i \in \Xi$, where Ξ is a convex compact subset of $\mathbb{R}_+^n \setminus \{0\}$. Clearly we can have information about the ‘true’ production function only on that set Ξ . We assume now that x_i are continuously distributed on Ξ .

(A.4) The input vectors x_i are iid random vectors having a continuous distribution whose support Ξ is a convex compact subset of $\mathbb{R}_+^n \setminus \{0\}$.

For some constant $c > 0$, let \mathcal{F}_c be the subset of \mathcal{F} formed by such functions $\phi \in \mathcal{F}$ that every supergradient $\nabla\phi(x)$ satisfies $\|\nabla\phi(x)\| \leq c$ for all $x \in \Xi$. This set \mathcal{F}_c is closed with respect to the sup-norm

$$\|\phi\| := \sup_{x \in \Xi} |\phi(x)| \quad (18)$$

Moreover, we have by the mean value theorem that every function $\phi \in \mathcal{F}_c$ is Lipschitz continuous modulus c , and hence it follows by the Arzela–Ascoli theorem (e.g., Billingsley, 1999, p. 81) that the set \mathcal{F}_c is compact with respect to this sup-norm. We assume that the constant c is large enough such that the true function f belongs to the set \mathcal{F}_c . Since a function $\phi \in \mathcal{F}$ is concave on \mathbb{R}_+^n , its supergradients are uniformly bounded on any compact subset of \mathbb{R}_{++}^n . Yet it may happen that these supergradients are unbounded at points arbitrary close to the boundary of the set \mathbb{R}_+^n . So by saying that $f \in \mathcal{F}_c$ we assume that this does not happen for the true function f .

Let \hat{f}_N be an optimal solution of the (restricted) problem:

$$\text{Min}_{\phi \in \mathcal{F}_c} \sum_{i=1}^N \theta \left(\frac{y_i}{\phi(x_i)} \right) \quad (19)$$

In view of Lemma A.1 the following result concerning consistency of the estimators \hat{f}_N should not be surprising. Such consistency results go back to the pioneering work of Wald (1949), and were discussed extensively in the statistics literature. We quickly outline its proof for the sake of completeness.

Theorem A.2 Suppose that assumptions (A.1)–(A.4) hold and $f \in \mathcal{F}_c$. Then \hat{f}_N converges, with respect to the sup-norm (5.13), w.p.1 as $N \rightarrow \infty$ to the true function f .

Proof For $\phi \in \mathcal{F}$ consider the functional

$$L_N(\phi) := N^{-1} \sum_{i=1}^N \theta \left(\frac{y_i}{\phi(x_i)} \right)$$

Note that $\hat{f}_N \in \arg \min_{\phi \in \mathcal{F}_c} L_N(\phi)$. By the law of large numbers (LLN) we have that, for any fixed $\phi \in \mathcal{F}$, $L_N(\phi)$ converges w.p.1 as $N \rightarrow \infty$ to the expectation

$$\ell(\phi) := \mathbb{E} \left[\theta \left(\eta \frac{f(X)}{\phi(X)} \right) \right] = \mathbb{E}_X \left\{ \mathbb{E} \left[\theta \left(\eta \frac{f(X)}{\phi(X)} \right) | X \right] \right\} \quad (20)$$

where X is a random variable distributed according to the distribution of the input vectors x_i . By applying Lemma A.1 we obtain that, for a given X , the minimum of $\mathbb{E} \left[\theta \left(\eta \frac{f(X)}{\phi(X)} \right) | X \right]$ over positive values of $\phi(X)$ is attained at $\phi(X) = f(X)$. Consequently, because by assumption (A.4) the support of X is Ξ , we obtain that $\ell(\phi)$ attains its minimum, over the set \mathcal{F}_c , at $\phi = f$ and this minimizer is unique. The remainder of the proof is rather standard. Since \mathcal{F}_c is compact the convergence (w.p.1) of $L_N(\cdot)$ to $\ell(\cdot)$ is uniform on \mathcal{F}_c , and the convergence of \hat{f}_N to f follows by compactness arguments. We refer to Newey and McFadden (1994) and Matzkin (1994), for example, for a discussion of such consistency results. \square

The assumption that the input sequence x_i is iid (assumption (A.4)), in the above proof, was used to justify the application of the LLN to obtain convergence w.p.1 of $L_N(\phi)$ to $\ell(\phi)$. This assumption can be relaxed in several ways. Suppose, for instance, that x_i are viewed now as forming a *deterministic* sequence. Then we can simply postulate that $L_N(\phi)$ converges w.p.1 to function $\ell(\phi)$ defined as

$$\ell(\phi) := \int_{\Xi} \mathbb{E} \left[\theta \left(\eta \frac{f(x)}{\phi(x)} \right) \right] h(x) dx \tag{21}$$

where $h : \Xi \rightarrow \mathbb{R}_{++}$ is a density function. This is a form of LLN (recall that the sequence η_i of the error terms is assumed to be iid) with density $h(x)$ representing distribution of x_i over the set Ξ . By replacing assumption (A.4) with this assumption we can proceed as in the proof above to show consistency of the ML estimators.

APPENDIX B

In this appendix we describe procedures for calculating the marginal rate of technical substitution (RTS) and the elasticity of substitution (ES) used in our numerical analyses.

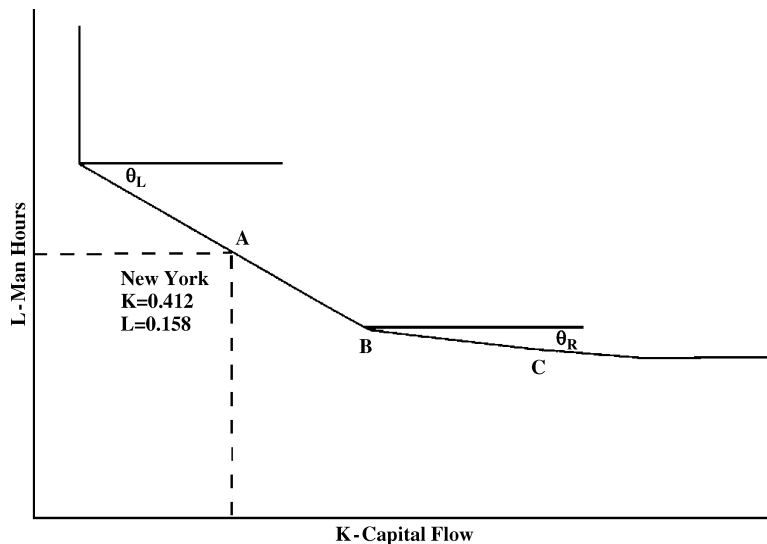


Figure 6. The level set of New York (CENP)

We begin with the marginal rate of technical substitution. Figure 6 depicts a typical level set, or isoquant, $L(y^*)$ in two dimensions, taken from an empirical application in Section 4. First consider point B, a vertex of the level set. Note that the slopes of the line segments $[A, B]$ and $[B, C]$, S_L and S_R , respectively, equal the tangent of the angles depicted, i.e., $S_L = \tan(\theta_L)$ and $S_R = \tan(\theta_R)$. Since B lies at the intersection of these two line segments, its ‘average angle’ is $\theta_{\text{avg}} := 1/2(\theta_L + \theta_R)$. The approximation we have devised for the RTS at B is given by $\tan(\theta_{\text{avg}})$. As for the extreme vertex located left of A, its $\theta_L = -\pi/2$ and for the extreme vertex located right of C, its $\theta_R = 0$. Finally, if the level set has only a single vertex, then its $\theta_L = -\pi/2$ and its $\theta_R = 0$.

Now consider a point $x = \lambda x_L + (1 - \lambda)x_R$ that lies in the interior of a line segment joining two vertices x_L and x_R whose corresponding average angles are θ_{x_L} and θ_{x_R} , respectively. In this case we shall define its RTS as $\tan(\lambda\theta_{x_L} + (1 - \lambda)\theta_{x_R})$. This approximation generates a continuous function, differentiable everywhere except at the vertices of the corresponding level sets.

Next, we consider the elasticity of substitution. Consider first a point x that is not a vertex. Let x_L and x_R denote the left and right neighboring vertices so that $x = \lambda x_L + (1 - \lambda)x_R$ for some $\lambda \in (0, 1)$. Observe that

$$\rho = \frac{x_2}{x_1} = \frac{\lambda x_{L2} + (1 - \lambda)x_{R2}}{\lambda x_{L1} + (1 - \lambda)x_{R1}} \quad (22)$$

so that

$$\lambda(\rho) = \frac{\rho x_{R1} - x_{R2}}{(x_{L2} - x_{R2}) - \rho(x_{L1} - x_{R1})} \quad (23)$$

By definition, RTS at x is given by $\text{RTS}(\rho) := \tan(\lambda(\rho)\theta_L + (1 - \lambda(\rho))\theta_R)$. Consequently, ES at x is given by

$$\text{ES}(x) := \frac{\rho}{\text{RTS}(\rho)} \times \text{RTS}'(\rho) = \frac{\rho}{\text{RTS}(\rho)} \times \frac{(\theta_L - \theta_R)\lambda'(\rho)}{\cos^2(\lambda(\rho)\theta_L + (1 - \lambda(\rho))\theta_R)} \quad (24)$$

where

$$\lambda'(\rho) = \frac{x_{R1}(x_{L2} - x_{R2}) - (x_{L1} - x_{R1})x_{R2}}{((x_{L2} - x_{R2}) - \rho(x_{L1} - x_{R1}))^2} \quad (25)$$

Now consider a vertex like B in Figure 6. Let $\text{ES}_L(0)$ denote its elasticity when it is viewed as the right endpoint of the line segment $[A, B]$, and let $\text{ES}_R(1)$ denote its elasticity when it is viewed as the left endpoint of the line segment $[B, C]$. We set its elasticity to be the $1/2(\text{ES}_L(0) + \text{ES}_R(1))$, the average of the two elasticities.

REFERENCES

- Abrevaya J, Jiang W. 2002. A simplex statistic for testing joint curvature. <http://www.mgmt.purdue.edu/faculty/abrevaya/simplex.pdf> [13 May 2006].
- Afriat SN. 1967. The construction of a utility function from expenditure data. *International Economic Review* **8**: 67–77.
- Afriat SN. 1971. The output limit function in general and convex programming and the theory of production. *Econometrica* **39**: 309–339.
- Afriat SN. 1972. Efficiency estimation of production functions. *International Economic Review* **13**: 568–598.

- Andrews DWK, Buchinsky M. 2001. A 3 step method for choosing the number of bootstrap repetitions. *Econometrica* **68**: 23–51.
- Arrow KJ, Chenery HB, Minhas BS, Solow RM. 1961. Capital–labor substitution and economic efficiency. *Review of Economics and Statistics* **43**: 225–250.
- Banker RD, Maundiratta A. 1992. Maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis* **3**: 401–415.
- Bazaraa MS, Shetty CM, Sherali H. 1993. *Nonlinear Programming: Theory and Algorithms* (2nd edn). Wiley: New York.
- Beenstock M. 1997. Business sector production in the short and long-run in Israel. *Journal of Productivity Analysis* **8**: 53–70.
- Ben-Tal A, Charnes A, Teboulle M. 1989. Entropic means. *Journal of Mathematical Analysis and Applications* **138**: 537–557.
- Billingsley P. 1999. *Convergence of Probability Measures* (2nd edn). Wiley: New York.
- Christensen LR, Jorgenson DW, Lau LJ. 1973. Transcendental logarithmic production frontiers. *Review of Economics and Statistics* **55**: 28–45.
- Cobb CW, Douglas PC. 1928. A theory of production. *American Economic Review, Supplement* **18**: 139–165.
- Csiszar I. 1967. Information type measurements of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **2**: 299–318.
- Doveh E, Shapiro A, Feigin PD. 2002. Testing of monotonicity in regression models. *Journal of Statistical Planning and Inference* **107**: 2289–2306.
- Gleser LJ, Moore DS. 1983. The effect of dependence in chi-squared and empiric distribution tests of fit. *Annals of Statistics* **11**: 1100–1108.
- Hall P, Huang L. 2001. Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics* **29**: 624–647.
- Hanoch G, Rothschild M. 1972. Testing the assumptions of production theory: a nonparametric approach. *Journal of Political Economy* **80**: 256–275.
- Härdle W. 1990. *Applied Nonparametric Regression*. Cambridge University Press.
- Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models*. Chapman & Hall: London.
- Lau LJ. 1986. Flexible functional forms. In *Handbook of Econometrics*, Vol. 3, Griliches Z, Intriligator M (eds). North-Holland: Amsterdam.
- Manski CF. 1995. *Identification Problems in the Social Sciences*. Harvard University Press: Boston, MA.
- Marschak J, Andrews W. 1944. Random simultaneous equations and the theory of production. *Econometrica* **12**: 143–153.
- Matzkin RL. 1991. Semiparametric estimation of monotone concave utility functions for polychotomous choice models. *Econometrica* **59**: 1351–1327.
- Matzkin RL. 1993. Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics* **58**: 137–168.
- Matzkin RL. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of Econometrics*, Vol. IV, Engle RF, McFadden DL (eds). North-Holland: Amsterdam.
- Matzkin RL. 1999. Computation of nonparametric concavity restricted estimators. Mimeo.
- Newey WK, MacFadden DL. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Vol. IV, Engle RF, McFadden DL (eds). North-Holland: Amsterdam.
- Robertson T, Wright FT, Dykstra RL. 1988. *Order Restricted Statistical Inference*. Wiley: Chichester.
- Varian HR. 1982. The nonparametric approach to demand analysis. *Econometrica* **50**: 945–973.
- Varian HR. 1983. Nonparametric tests of consumer behavior. *Review of Economic Studies* **50**: 99–110.
- Varian HR. 1984. The nonparametric approach to production analysis. *Econometrica* **52**: 579–597.
- Varian HR. 1985. Nonparametric analysis of optimizing behavior with measurement error. *Journal of Econometrics* **30**: 445–458.
- Wald A. 1949. Note on the consistency of the maximum likelihood estimates. *Annals of Mathematical Statistics* **20**: 595–601.
- Yatchew AJ, Bos L. 1997. Nonparametric least squares regression and testing in economic models. *Journal of Quantitative Economics* **13**: 81–131.
- Zellner A, Ryu H. 1998. Alternative functional forms for production, cost and returns to scale functions. *Journal of Applied Econometrics* **13**: 101–127.