Econometrics Journal (2008), volume **11**, pp. 308–325.
doi: 10.1111/j.1368-423X.2008.00239.x

Representation theorem for convex nonparametric least squares

TIMO KUOSMANEN

Economic Research Unit, MTT Agrifood Research, Finland.
E-mail: Timo.Kuosmanen@mtt.fi

First version received: May 2007; final version accepted: January 2008

Summary We examine a nonparametric least-squares regression model that endogenously selects the functional form of the regression function from the family of continuous, monotonic increasing and globally concave functions that can be nondifferentiable. We show that this family of functions can be characterized without a loss of generality by a subset of continuous, piece-wise linear functions whose intercept and slope coefficients are constrained to satisfy the required monotonicity and concavity conditions. This representation theorem is useful at least in three respects. First, it enables us to derive an explicit representation for the regression function, which can be used for assessing marginal properties and for the purposes of forecasting and ex post economic modelling. Second, it enables us to transform the infinite dimensional regression problem into a tractable quadratic programming (QP) form, which can be solved by standard QP algorithms and solver software. Importantly, the QP formulation applies to the general multiple regression setting. Third, an operational computational procedure enables us to apply bootstrap techniques to draw statistical inference.

Keywords: *Concavity*, *Convexity*, *Curve fitting*, *Linear splines*, *Local linear approximation*, *Nonparametric methods*, *Regression analysis*.

1. INTRODUCTION

Nonparametric regression techniques that avoid strong prior assumptions about the functional form are attracting increasing attention in econometrics. Nonparametric least squares subject to continuity, monotonicity and concavity constraints [henceforth referred to as Convex Nonparametric Least Squares (CNLS)] are one of the oldest approaches, dating back to the seminal work by Hildreth (1954). This method draws its power from the shape constraints that coincide with the standard regularity conditions of the microeconomic theory (see e.g. Varian, 1982, 1984). In contrast to the kernel regression and spline smoothing techniques, CNLS does not require specification of a smoothing parameter. Thus, CNLS circumvents the fundamental bias-variance tradeoff (see e.g. Yatchew, 2003, for discussion) associated with most other nonparametric regression techniques.

Earlier work on CNLS has mainly focused on the statistical properties, and the essential aspects of the CNLS estimators are nowadays well understood. The maximum-likelihood interpretation of CNLS was already noted by Hildreth (1954), and Hanson and Pledger (1976) have proved its consistency. More recently, Nemirovskii et al. (1985), Mammen (1991) and Mammen and Thomas-Agnan (1999) have shown that CNLS achieves the standard nonparametric rate of

convergence $O_P(n^{-1/(2+m)})$ where n is the number of observations and m is the number of regressors. Imposing further smoothness assumptions or derivative bounds can improve the rate of convergence and alleviate the curse of dimensionality (see e.g. Mammen, 1991, Yatchew, 1998, Mammen and Thomas-Agnan, 1999, and Yatchew and Härdle, in press). Groeneboom et al. (2001) have derived the asymptotic distribution of the univariate CNLS estimator at a fixed point.

Despite the attractive theoretical properties, empirical applications of CNLS remain scarce. There are many factors restricting the diffusion of CNLS to econometrics. In our view, three major barriers are

(1) the lack of explicit regression function,
(2) computational complexity and
(3) difficulty of statistical inference.

The lack of a tractable closed-form expression for the CNLS regression function presents a clear disadvantage relative to the parametric methods and some other nonparametric techniques such as kernel smoothing. Indeed, economists are often interested in the marginal properties and elasticities of the function, which cannot be assessed based on a discrete set of fitted values provided by CNLS. Moreover, a simple closed-form expression of the regression function is necessary for using the regression results in ex post economic modelling (e.g. using estimated utility and production functions in a computable general equilibrium model).

The computational complexity of CNLS is due to the fact that the functional form of the regression function is not assumed a priori, but it is endogenously selected from an infinitely large family of continuous, monotonic increasing and concave functions. Efficient computational algorithms have been developed by Wu (1982), Fraser and Massam (1989), Goldman and Ruud (1995), Ruud (1996) and Meyer (1999), but the implementation of these procedures requires considerable programming skills. More importantly, most existing computational procedures are restricted to the single regressor case where the observations can be sorted according to the explanatory variable. These algorithms cannot be generalized (even in principle) to the multiple regressor setting involving a vector of regressors.

In principle, conventional methods of statistical inference could be adapted to the context of CNLS. However, the degrees of freedom depend on the number of different hyperplane segments or the number of observations projected to a given segment (see Meyer, 2003, 2006, for discussion). Moreover, the segments are endogenously determined in the model, and the coefficients of the segments may not be unique in the multiple regression setting. For these reasons, the bootstrap approach appears to be the most promising tool for statistical inferences (see e.g. Efron, 1979, 1982, and Efron and Tibshirani, 1993). However, implementing computationally intensive bootstrap simulations requires a fast, tractable algorithm for computing the estimator.

This paper presents a representation theorem that helps us to overcome (or at least lower) each of these three barriers. Firstly, drawing insight from the celebrated Afriat's Theorem, we derive an explicit representor function which can be used for assessing marginal properties and for the purposes of forecasting and ex post economic modelling. The representor function is a simple piece-wise linear function that is easy to compute given the coefficients estimated by CNLS. Secondly, the representation theorem is useful from the computational point of view: it enables us to transform the infinite dimensional CNLS problem into a tractable quadratic programming (QP) form, which can be solved by standard QP algorithms and solver software. Importantly, the QP formulation applies to the general multiple regression setting. Thirdly, existence of a tractable computational procedure enables one to apply computationally intensive bootstrap or

Monte Carlo simulations to draw statistical inference or assess the small sample performance of the estimator, respectively. Finally, in addition to tackling these three barriers, we point out a number of interesting links between CNLS and parallel developments in the literature.

The rest of the paper is organized as follows. Section 2 presents our main result. Section 3 applies the result to formulate the infinite dimensional CNLS problem as a finite dimensional QP problem. Section 4 derives an explicit representor function that provides the first-order approximation for any arbitrary regression function in the neighbourhood of observed points. Section 5 illustrates the potential of the method by means of Monte Carlo simulations. Section 6 presents a concluding discussion and points some directions for future research. In the interest of readability, all formal proofs of mathematical theorems are presented in Appendix A. A GAMS code for computing the CNLS regression is presented in Appendix B.

2. THE REPRESENTATION THEOREM

Consider the canonical multiple regression model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, i = 1, \ldots, n, \tag{2.1}$$

where y_i is the dependent variable, f is an unknown regression function to be estimated, $\mathbf{x}_i \in \mathbb{R}^m$ is the vector of explanatory variables and ε_i is the idiosyncratic error term. Errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ are assumed to be uncorrelated random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty \forall i = 1, \ldots, n$ (i.e. the Gauss–Markov conditions). The data set of n observations is denoted by $(\mathbf{X},\mathbf{y})$, with $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{m \times n}$.

In contrast to the linear and nonlinear parametric approaches, we assume no particular functional form for f a priori. Instead, we impose a more general condition that f belongs to the set of continuous, monotonic increasing and globally concave functions denoted by

$$F_2 = \left\{ f : \mathbb{R}^m \to \mathbb{R} \;\middle|\; \begin{array}{l} \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^m : \mathbf{x} \geq \mathbf{x}' \Rightarrow f(\mathbf{x}) \geq f(\mathbf{x}'); \\ \forall \mathbf{x}', \mathbf{x}'' \in \mathbb{R}^m : \mathbf{x} = \lambda \mathbf{x}' + (1-\lambda)\mathbf{x}'', \lambda \in [0,1] \Rightarrow f(\mathbf{x}) \geq \lambda f(\mathbf{x}') + (1-\lambda) f(\mathbf{x}'') \end{array} \right\}. \tag{2.2}$$

The rationale behind the continuity, monotonicity and concavity postulates lies in their central role in the microeconomic theory (see e.g. Varian, 1982, 1984).

The CNLS problem is to find $f \in F_2$ that minimizes the sum of squares of the residuals, formally:

$$\min_f \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \quad s.t. f \in F_2. \tag{2.3}$$

In other words, the CNLS estimator of f is a monotonic increasing and concave function that minimizes the L_2-norm of the residuals. The CNLS problem (2.3) does not restrict beforehand to any particular functional form, but selects the best-fitting function f from the family F_2, which includes an infinite number of functions. This makes problem (2.3) generally hard to solve. Single regressor algorithms developed by Wu (1982), Fraser and Massam (1989) and Meyer (1999) require that the data are sorted in ascending order according to the scalar-valued regressor x. However, such a sorting is not possible in the general multiple regression setting where $\mathbf{x}$ is a vector.

The Sobolev least-squares models (e.g. Wahba, 1990) differ from CNLS in that functions f must be differentiable (smooth) at every point of its domain and the Sobolev norm of f is bounded from above. The Sobolev models that impose monotonicity and concavity (e.g. Yatchew and Bos, 1997, and Yatchew and Härdle, in press) are hence constrained variants of (2.3). Inspired by that literature, we next try to identify a subset of representor functions $G{:}G \subset F_2$ such that replacing the constraint $f \in F_2$ by $f \in G$ does not influence the optimal solution of problem (2.3) but makes it easier to solve.

Consider the following family of piece-wise linear functions

$$G_2(\mathbf{X}) = \{g : \mathbb{R}^m \to \mathbb{R} \,|\, g(\mathbf{x}) = \min_{i \in \{1,\ldots,n\}} \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}; \tag{2.4}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0} \quad \forall i = 1, \ldots, n; \tag{2.5}$$

$$\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \quad \forall h, i = 1, \ldots, n\Big\}. \tag{2.6}$$

Clearly, functions $g \in G_2(\mathbf{X})$ are continuous, monotonic increasing and globally concave for any arbitrary $\mathbf{X}$. Hence $G_2 \subset F_2$. The following theorem shows that this class of functions can be used as representors when solving the infinite dimensional CNLS problem (2.3).

THEOREM 2.1. Given an arbitrary finite real-valued data ($\boldsymbol{X}$,$\boldsymbol{y}$), denote the optimal solution to the CNLS problem (2.3) by s_f^2 and let

$$s_g^2 \equiv \min_g \sum_{i=1}^{n} (y_i - g(\mathbf{x}_i))^2 \quad s.t.\, g \in G_2(\mathbf{X}). \tag{2.7}$$

Then $s_f^2 = s_g^2$.

This result augments the representation theorems of the Sobolev least squares (e.g. Yatchew and Bos, 1997) to the nonsmooth CNLS setting. A number of parallel results are known in the literature. In the microeconomic theory, the celebrated Afriat's Theorem relates continuous, monotonic, concave utility functions with piece-wise linear representors in a directly analogous way (Afriat, 1967, and Varian, 1982). In the productive efficiency literature, Banker and Maindiratta (1992) have applied the Afriat inequalities in the maximum-likelihood estimation of frontier production functions perturbed by skewed, non-Gaussian error terms. In the present context of nonparametric regression, the possibility to use the Afriat inequalities to model concavity/convexity constraints has been briefly suggested by Matzkin (1994, 1999) and Yatchew (1998); Theorem 2.1 confirms and further elaborates these conjectures. In the context of limited dependent variable models, Matzkin (1991, 1992) has employed the Afriat inequalities to develop consistent semi- and nonparametric estimators for the consumer's utility function. Finally, Mammen (1991) has derived a similar theorem for a class of nonparametric regression functions constrained by qualitative shape restrictions, showing that these functions can be represented by a class of piece-wise monotonic or piece-wise concave/convex splines.

The link between CNLS and spline smoothing becomes evident if we interpret the piece-wise linear representors $g \in G_2(\mathbf{X})$ as linear spline functions. In contrast to the spline functions, however, here the partition to the linear segments is not fixed a priori. Indeed, the number and the location of the segments are here endogenously determined to maximize the empirical fit. Theorem 2.1 implies that the CNLS problem (2.3) can be equivalently stated as a linear spline

smoothing problem where the knots (i.e. the vertices of the hyperplane segments) are optimally selected to minimize the sum of squares of residuals subject to the monotonicity and concavity constraints for the spline function. It is worth to emphasize that the knots do not generally coincide with the observed points, but occur typically somewhere between them (see Figure 1 below for a graphical illustration). Moreover, the number of knots is usually a small fraction of n.

Theorem 2.1 extends in a straightforward fashion to globally convex and/or monotonic decreasing functions. In the case of a convex regression function, the signs of the inequality constraints (2.6) should be reversed. Similarly, a monotonic decreasing function is obtained by reversing the signs of the inequalities (2.5). One could easily impose further assumptions about linear homogeneity. If function f is known to be homogenous of degree one (e.g. if f is a production function exhibiting constant returns to scale or an expected utility function exhibiting risk neutrality), we may simply impose an additional constraint $\alpha_i = 0$ (or delete α_i altogether) in (2.4)–(2.6). This will guarantee that functions g pass through the origin. On the other hand, the monotonicity constraints (2.5) could be relaxed to estimate (inverse) U-shaped curves. However, removing the concavity constraints (2.6) does not directly lead us to the isotonic regression model considered by Barlow et al. (1972) and Sasabuchi et al. (1983). Establishing formal links between CNLS and isotonic regression formulations and exploring the intermediate cases of quasi-concave/convex regression are left as a challenge for future research.

3. QUADRATIC PROGRAMMING FORMULATION

Theorem 2.1 is important from the computational point of view. It enables us to transform the infinite dimensional problem (2.3) into the finite dimensional QP problem (2.7). This QP formulation can be expressed more intuitively as

$$\begin{aligned} \min_{\varepsilon,\alpha,\beta} & \sum_{i=1}^{n} \varepsilon_i^2 \\ & y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i \\ & \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i \quad \forall h, i = 1, \ldots, n \\ & \boldsymbol{\beta}_i \geq 0 \, \forall i = 1, \ldots, n. \end{aligned} \tag{3.1}$$

The first constraint of (3.1) can be interpreted as the regression equation: note that coefficients $\alpha_i, \boldsymbol{\beta}_i$ are specific to each observation $i : i = 1, \ldots, n$. The second constraint is a system of Afriat inequalities that guarantee concavity [equivalent to (2.6)]. The Afriat inequalities are the key to modelling concavity constraints in the general multiple regressor setting. The third constraint ensures monotonicity [equivalent to (2.5)].

The QP problems represent the simplest thinkable class of nonlinear optimization problems; many sophisticated algorithms and powerful solvers are nowadays available for such problems.[1]

[1]QP is a standard class of problems within nonlinear programming (NLP). The quadratic objective function implies that the first-order conditions are linear. Hence, the QP problems are amenable to the standard simplex and interior point algorithms developed for linear programming. A variety of commercial and sharewere solver software are available for solving QP problems. High-performance QP solvers include, e.g., CPLEX, LINDO, MOSEK and QPOPT, but also general NLS solvers such as MINOS and BQPD can handle QP problems. Most solvers can be used integrated with standard mathematical modelling systems/languages such as GAMS, Gauss, Mathematica and Matlab.

The QP formulation of the CNLS problem presents a clear advantage to the earlier computational algorithms that only apply in the single regression case (e.g. Wu, 1982, Fraser and Massam, 1989, and Meyer, 1999).

It is worth to note the structural similarity between problem (3.1) and the varying coefficient (or random parameters) regression models (e.g. Fan and Zhang, 1999, and Greene, 2005). The varying coefficient models assume a conditional linear structure that allows the coefficients of the linear regression function to differ across (groups of) observations, similar to (3.1). Interestingly, our representation theorem shows that the varying coefficient approach (conventionally used for estimating n different regression functions of the same a priori specified functional form) can be used for estimating n tangent hyper-planes to a single, unspecified regression function.

The piece-wise linear structure of CNLS also resembles the nonparametric data envelopment analysis (DEA) frontiers (compare with Banker and Maindiratta, 1992). The key difference between CNLS and DEA concerns the treatment of residuals ε_i. In DEA, the residual term is interpreted as deterministic inefficiency that can only take negative values. The standard variable returns to scale DEA model is obtained as a special case of (3.1) if the residuals are constrained to be non-positive [i.e. one adds constraint $\varepsilon_i \leq 0 \ \forall i = 1, \ldots, n$ to problem (3.1); see Kuosmanen, 2006 for details].

4. DERIVING AN EXPLICIT REPRESENTOR FUNCTION

This section takes a more detailed view on the representor functions g. Interestingly, the solution to a complex problem need not itself be very complex: Theorem 2.1 shows that the infinite dimensional optimization problem (2.3) always has an optimal solution that takes the form of a simple piece-wise linear function. Given the estimated coefficients $(\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i)$ from (3.1), we can construct the following explicit *representor function*

$$\hat{g}(\mathbf{x}) \equiv \min_{i \in \{1,\ldots,n\}} \left\{ \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \mathbf{x} \right\}. \tag{4.1}$$

In principle, function $\hat{g}$ consists of n hyperplane segments. In practice, however, the estimated coefficients $(\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i)$ are clustered to a relatively small number of alternative values: the number of different hyperplane segments is usually much lower than n (see Section 5 for some simulation evidence). When the number of different segments embedded in (4.1) is small, the values of $\hat{g}$ are easy to enumerate. The simplicity of the representor is an appealing feature for its potential ex post uses in economic modelling. The use of $\hat{g}$ as an estimator of f is justified by the following result:

COROLLARY 4.1. Denote the set of functions that minimize the CNLS problem (2.3) by F_2^*. For any finite real-valued data $(\boldsymbol{X}, \boldsymbol{y})$, the function $\hat{g}$ defined by (4.1) and (3.1) is one of the optimal solutions to problem (2.3), that is, $\hat{g} \in F_2^*$.

The representor $\hat{g}$ and its coefficients $(\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i)$ have a compelling interpretation: vector $\hat{\boldsymbol{\beta}}_i$ can be interpreted as an estimator of the subgradient vector $\nabla f(\mathbf{x}_i)$, and equation $y = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \mathbf{x}$ is an estimator of the tangent hyperplane of f at point $\mathbf{x}_i$. In other words, function $\hat{g}$ provides a local first-order Taylor series approximation to any $f \in F_2^*$ in the neighbourhood of the observed

points $\mathbf{x}_i$.[2] This justifies the use of the representor $\hat{g}$ for forecasting the values of y not just at the observed points, but also at unobserved points in the neighbourhood of observations.

Coefficients $\hat{\boldsymbol{\beta}}_i$ can also be used for nonparametric estimation of the marginal properties and elasticities. We can calculate the rate of substitution between variables k and m at point $\mathbf{x}_i$ as

$$\frac{\partial \hat{g}(\mathbf{x}_i)\big/\partial x_k}{\partial \hat{g}(\mathbf{x}_i)\big/\partial x_m} = \frac{\hat{\beta}_{ik}}{\hat{\beta}_{im}}, \tag{4.2}$$

and further, the elasticity of substitution as

$$e_{k,m}(\mathbf{x}_i) = \frac{\hat{\beta}_{ik}}{\hat{\beta}_{im}} \cdot \frac{x_{im}}{x_{ik}}. \tag{4.3}$$

These substitution rates and elasticities are simple to compute given the estimated $\hat{\boldsymbol{\beta}}_i$ coefficients.

One should note that the optimal solution to problem (2.3) is not necessarily unique; there generally exists a family of alternate optima, denoted by F_2^*. The optimal solution to problem (3.1) need not be unique either, although the fitted values and most of the coefficients typically do have a unique solution. The set of alternative representor functions characterized by (3.1) and (4.1) form a subset of F_2^*. The lack of a unique solution might be seen as a serious problem of identification, but this does not render the CNLS model meaningless. In fact, it is possible to derive tight lower and upper bounds for the alternate optima within F_2^*. Specifically, functions $f \in F_2^*$ are bounded by the following piece-wise linear functions

$$\hat{g}_{\min}(\mathbf{x}) = \min_{\alpha\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^m} \left\{\alpha + \boldsymbol{\beta}' x \,\middle|\, \alpha + \boldsymbol{\beta}' x_i \geq \hat{y}_i \quad \forall i = 1,\ldots,n\right\}, \tag{4.4}$$

and

$$\hat{g}_{\max}(\mathbf{x}) = \max_{\phi\in\mathbb{R},\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^{m\times n}} \left\{\phi \,\middle|\, \phi \leq \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}\ \forall i; \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i = \hat{y}_i\ \forall i; \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_h \geq \hat{y}_h\ \forall h \neq i\right\}, \tag{4.5}$$

where

$$\hat{y}_i = \hat{g}(\mathbf{x}_i) = y_i - \hat{\varepsilon}_i, i = 1,\ldots,n, \tag{4.6}$$

denote the fitted values of the dependent variable.

THEOREM 4.1. For any finite real-valued data $(\boldsymbol{X}, \boldsymbol{y})$, function $\hat{g}_{\min}$ is the tightest possible lower bound for the family of functions F_2^*, and $\hat{g}_{\max}$ is the tightest possible upper bound for F_2^*. Specifically,

$$\hat{g}_{\min}(\mathbf{x}) = \min_f f(\mathbf{x})\, s.t. f \in F_2^* \tag{4.7}$$

and

$$\hat{g}_{\max}(\mathbf{x}) = \max_f f(\mathbf{x})\, s.t. f \in F_2^* \tag{4.8}$$

for all $\mathbf{x} \in \mathbb{R}^m$.

[2]In contrast to the flexible functional forms that can be interpreted as second-order Taylor approximations around a single, unknown expansion point, CNLS uses all n observations as expansion points for the local linear approximation.

Theorem 4.1 further highlights the role of the piece-wise linear functions in CNLS. The boundary functions $\hat{g}_{\min}$ and $\hat{g}_{\max}$ are analogous to the under- and overproduction functions derived by Varian (1984). The lower bound $\hat{g}_{\min}$ satisfies the maintained regularity properties. Moreover, since $\hat{g}_{\min}(\mathbf{x}_i) = \hat{y}_i \quad \forall i = 1, \ldots, n$, we have $\hat{g}_{\min} \in F_2^*$: there exists a unique piece-wise linear function that characterizes the lower boundary of family F_2^* for all $\mathbf{x} \in \mathbb{R}^m$. By contrast, the upper bound $\hat{g}_{\max}$ is not globally concave, and thus $\hat{g}_{\max} \notin F_2^*$. For any given $\mathbf{x} \in \mathbb{R}^m$, the upper bound $\hat{g}_{\max}(\mathbf{x})$ is achieved by some $f \in F_2^*$, but in general, no concave function is able to reach the upper bound of F_2^* at all points $\mathbf{x} \in \mathbb{R}^m$ simultaneously.

In small samples, the boundary functions satisfy inequalities $\hat{g}_{\min}(\mathbf{x}) \leq \hat{g}(\mathbf{x}) \leq \hat{g}_{\max}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^m$. The consistency result by Hanson and Pledger (1976) implies that $\hat{g}_{\min}(\mathbf{x}) - \hat{g}_{\max}(\mathbf{x}) \xrightarrow{p} 0$ as $n \to \infty$. Given a large enough density of observations within the observed range, the lower and upper bounds coincide with the representor $\hat{g}$.

5. MONTE CARLO SIMULATIONS

5.1. Single regression example

This section examines the CNLS representor function and the method as a whole by simple Monte Carlo simulations.[3] To gain intuition, we first illustrate the CNLS representor in the single regression case. Suppose the true regression function is of the form $f(x) = \ln(x) + 1$. We drew a random sample of 100 observations of the x values from $Uni[1,11]$, calculated the corresponding true $f(x_i)$ values and perturbed them by adding a random error term drawn independently from $N(0, 0.6^2)$. This gives the observed y_i values for the dependent variable. Figure 1 illustrates the observed sample by the scatter plot. Also the true function f (solid grey curve) is plotted in the figure.

We solved the QP problem (3.1) by using the GAMS software with Minos nonlinear programming (NLP) solver. The coefficient of determination was $R^2 = 0.795$. The optimal solution to (3.1) provides coefficients $\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i$, which were used for constructing the piece-wise linear representor $\hat{g}$, plotted in Figure 1. This function consists of six different line segments; the estimated $\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i$ coefficients were clustered to six different vectors in this example. Recall that, in contrast to the linear spline smoothing methods, the positions of the line segments are not fixed ex ante but the number and the length of the segments are endogenously determined within the model. As Figure 1 shows, function $\hat{g}$ (solid black curve) provides a good approximation of the true f throughout the observed range of x, not only at the observed points but also in their neighbourhood. To appreciate this result, we also fitted the log-linear Cobb–Douglas function with OLS (broken grey curve). As Figure 1 indicates, the Cobb–Douglas function proved too inflexible for capturing the shape of the true f in this example.

5.2. Multiple regression simulations

We next performed more systematic Monte Carlo simulations in the two-regressor setting, fixing the sample size at 100 as before. Three different specifications for the true regression function f were considered:

[3]For empirical applications, an interested reader is referred to working papers Kuosmanen (2006) and Kuosmanen and Kortelainen (2007) that apply CNLS to production frontier estimation in cross-sectional and panel settings.

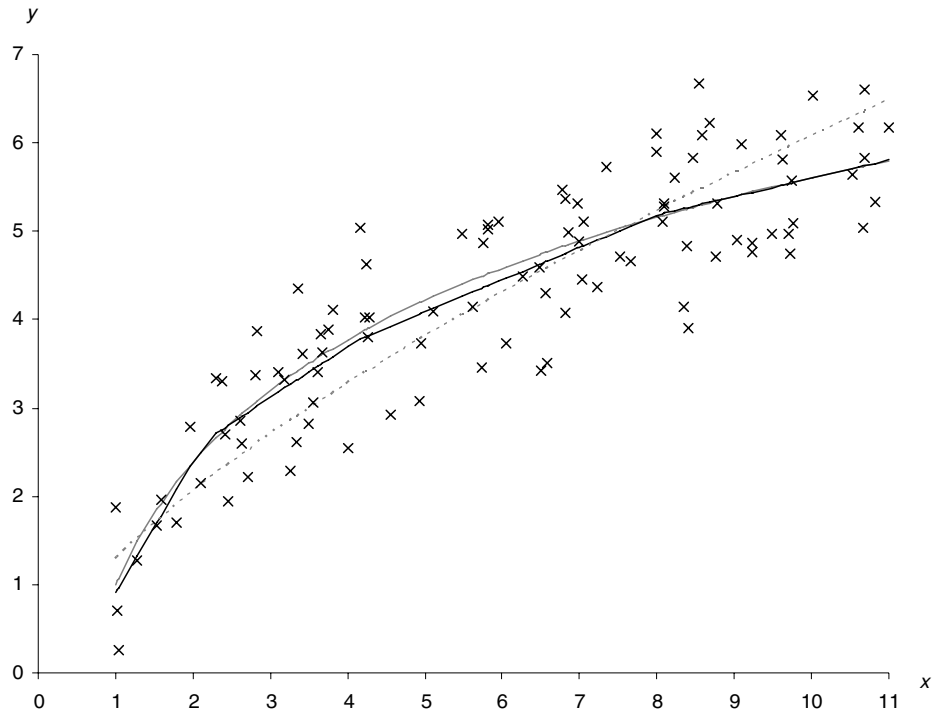

Figure 1. Illustration of $\hat{g}$ of the CNLS regression.

(1) Cobb–Douglas: $f^{\mathrm{CD}}(x_1, x_2) = x_1^{0.4} \cdot x_2^{0.5}$
(2) Generalized Leontief: $f^{\mathrm{GL}}(x_1, x_2) = (0.2x_1^{0.5} + 0.3x_2^{0.5} + 0.4x_1^{0.5} \cdot x_2^{0.5})^{0.9}$
(3) Piece-wise linear: $f^{\mathrm{PWL}}(x_1, x_2) = \min\{x_1 + 2x_2, 2x_1 + x_2, 0.5x_1 + x_2 + 225, x_1 + 0.5x_2 + 225\}$.

The values x_1 and x_2 were independently and randomly sampled from the uniform distribution *Uni*[100,200], and the true $f(x_1, x_2)$ were computed. Subsequently, random error terms drawn from $N(0, \sigma^2)$ were added to $f(x_1, x_2)$. Three different levels of standard deviation were considered: (A) low $\sigma = 2.5$, (B) medium $\sigma = 5$ and (C) high $\sigma = 10$. The resulting data sets perturbed by errors were treated as the observations. We computed 250 simulations for each of the nine alternative scenarios, referred to as (1A), (1B), . . . , (3C). The GAMS code for solving the CNLS model is presented in Appendix B.

First, we know that the CNLS representor $\hat{g}$ satisfies the regularity properties globally, but how useful the estimated $\hat{g}$ functions can be in practical economic modelling? The answer to this question obviously depends on the complexity of the functions. Table 1 sheds further light on this question by describing the number of hyperplane segments in each scenario. We find that typically about 20 hyper-plane segments suffice to characterize the CNLS function; in some scenarios the number was less than 10, in others, as high as 50. Relatively small number of segments is desirable for analytical and computational convenience as well as for narrowing the gap between lower and upper bounds. Somewhat surprisingly, CNLS uses large numbers of segments in Scenario 3 where

Table 1. CNLS representor: the number of different hyperplane segments in each scenarios.

True function	Error variance	Signal-to noise ratio∗	Average no. of segments	St.Dev. no. of segm.	Min no. of segm.	Max no. of segm.	% observations with $\hat{\beta}_m = 0$
Cobb–	$\sigma = 2.5$	4.5	23.1	4.9	9	35	5.3
Douglas	$\sigma = 5$	2.25	20.8	4.7	6	33	8.6
	$\sigma = 10$	1.13	18.7	4.7	6	30	7.6
Generalized	$\sigma = 2.5$	2	20.5	4.8	6	34	8.9
Leontief	$\sigma = 5$	1	18.7	4.9	6	32	14.7
	$\sigma = 10$	0.5	16.8	4.4	7	27	25.3
Piece-wise	$\sigma = 2.5$	20	36.9	4.4	25	49	2.7
linear	$\sigma = 5$	10	37.7	4.9	24	51	4.8
	$\sigma = 10$	5	33.7	4.9	19	48	8.0

∗Measured by the expected standard deviation of $f(x_1, x_2)$ values divided by σ.

the true piece-wise linear function consists of only four linear segments. This seems to be due to the fact that the number of hyperplane segments is positively correlated with the signal-to-noise ratio: the noisier the data, the smaller the number of segments.

The right-most column of Table 1 reports the percentage of observations for which any of the elements in vector $\hat{\boldsymbol{\beta}}_i$ is zero. Recall that these slope coefficients determine the substitution properties. While the zero values are consistent with the regularity conditions, the economic interpretation of the regression becomes odd if there are many zero substitution rates. In these simulations, the percentages of zero coefficients are relatively small.

The average number of observations per hyperplane segment was four; in Scenarios 1 and 2 this average exceeded five. This means that the bounds $\hat{g}_{\min}$ and $\hat{g}_{\max}$ coincided with $\hat{g}$ for very large proportions of the curve. The percentage of observations i such that $\hat{g}_{\min}(\mathbf{x}_i) = \hat{g}_{\max}(\mathbf{x}_i)$ varied between 70 and 95 per scenario, with the mean value of 87.4 across all scenarios. That is, the min and max bounds coincide for almost 90 per cent area of the estimated surfaces. The deviations typically occurred near the boundaries of the observed range of $\mathbf{x}$.

Consider next the empirical fit. Table 2 reports the coefficient of determination (R^2), log-likelihood ($\ln L$) and the mean-squared error (MSE) statistics for each scenario. To put the performance of CNLS in a perspective, we also estimated the Cobb–Douglas and translog regression functions using OLS.

The CNLS always gave the highest R^2 and log-likelihood values as expected. The difference is notable especially in Scenarios 2C and 3A. Moreover, we see that the empirical fit and the MSE values tend to deteriorate as the error variance increases (the only exception is the MSE of the translog regression that decreased in Scenario 3.

In Scenario 1, we see that the correctly specified Cobb–Douglas model yields the lowest MSE. The empirical fit of the more flexible CNLS and translog models becomes 'too good' in this scenario; the error variance is underestimated. The translog specification performs somewhat better than the CNLS in Scenario 1, but the difference is relatively small. In Scenarios 2A and 2B (the standard deviation of 2.5 and 5), the translog regression yields the lowest MSE, which is hardly surprising since both translog and generalized Leontief belong to the family of flexible function

Table 2. Goodness of fit: average of 250 simulations (standard deviation in parentheses).

True function	Std. dev.	CNLS estim. R^2	CNLS estim. LnL	CNLS estim. MSE	Cobb–Douglas estim. R^2	Cobb–Douglas estim. ln$L$	Cobb–Douglas estim. MSE	Translog estim. R^2	Translog estim. ln$L$	Translog estim. MSE
Cobb–	$\sigma = 2.5$	0.960	−179	0.79	0.954	−187	0.22	0.954	−186	0.41
Douglas		(0.008)	(8.54)	(0.34)	(0.009)	(6.83)	(0.16)	(0.009)	(7.85)	(0.23)
	$\sigma = 5$	0.856	−250	2.71	0.838	−256	0.89	0.841	−255	1.68
		(0.029)	(8.34)	(1.23)	(0.029)	(6.83)	(0.64)	(0.030)	(7.86)	(0.92)
	$\sigma = 10$	0.606	−321	9.48	0.561	−326	3.99	0.575	−325	7.09
		(0.068)	(8.16)	(4.47)	(0.070)	(7.54)	(2.97)	(0.070)	(7.88)	(3.81)
Gener.	$\sigma = 2.5$	0.830	−181	0.66	0.813	−187	7.58	0.809	−187	0.23
Leontief		(0.034)	(8.31)	(0.30)	(0.031)	(6.83)	(1.04)	(0.035)	(7.52)	(0.16)
	$\sigma = 5$	0.562	−252	2.33	0.530	−256	38.46	0.518	−256	1.00
		(0.073)	(8.13)	(1.10)	(0.064)	(6.86)	(8.97)	(0.074)	(7.54)	(0.75)
	$\sigma = 10$	0.281	−322	8.19	0.053	−557	46821	0.045	−335	111
		(0.084)	(7.97)	(4.02)	(0.095)	(143)	(40176)	(0.259)	(12.8)	(132)
Piece-	$\sigma = 2.5$	0.998	−171	1.65	0.973	−305	58.94	0.973	−305	59.18
wise		(0.000)	(9.66)	(0.50)	(0.005)	(6.33)	(8.04)	(0.004)	(6.08)	(7.16)
linear	$\sigma = 5$	0.992	−242	5.58	0.965	−318	59.79 (8.20)	0.985	−276	15.07
		(0.001)	(9.40)	(1.80)	(0.006)	(6.32)		(0.003)	(7.71)	(2.27)
	$\sigma = 10$	0.969	−314	17.86	0.963	−350	62.83	0.957	−330	20.00
		(0.007)	(9.08)	(6.20)	(0.005)	(6.78)	(9.00)	(0.01)	(7.89)	(4.36)

forms. The MSE statistics of the CNLS come very close to those of translog in these two scenarios, while Cobb–Douglas has notably higher MSE. When standard deviation is increased to 10 in Scenario 2C, the CNLS still provides reasonably accurate estimates, but the performance of the Cobb–Douglass and translog regressions is catastrophic. In Scenario 3, the CNLS has the lowest MSE throughout, as expected. The Cobb–Douglas specification performs poorly throughout all sub-scenarios, while the translog performs relatively well when the standard deviation of error increases to 10. Overall, the MSE statistics suggest the CNLS provides more robust performance than the two parametric candidates considered. Increasing the number of observations would further improve the accuracy of CNLS, whereas increasing the number of explanatory variables would likely favour OLS.

We conclude by emphasizing that consistency with the regularity properties implied by the economic theory is often a more important criterion than the empirical fit. In this respect, CNLS will always satisfy monotonicity and concavity by construction. The Cobb–Douglas function satisfies monotonicity but can violate concavity, while translog can violate both. Table 3 reports the frequencies of violations in each scenario. Our simulations suggest that the parametric regression models are surprisingly likely to violate the regularity conditions even when the true functions satisfy the properties and the empirical fit is reasonably good. For example, in Scenario 2C, 80 per cent of the estimated Cobb–Douglas functions were convex although the true underlying function

Table 3. Violations of concavity and monotonicity (per cent of simulations).

True function	Error variance	Cobb–Douglas concavity	Translog concavity	Translog monotonicity
Cobb–Douglas	$\sigma = 2.5$	0	52.6	0.000
	$\sigma = 5$	0.4	44.3	0.016
	$\sigma = 10$	23.2	41.7	1.26
Generalized Leontief	$\sigma = 2.5$	0	0	0
	$\sigma = 5$	14.4	0	0
	$\sigma = 10$	80	0	0
Piece-wise linear	$\sigma = 2.5$	0	0	0
	$\sigma = 5$	0	100	0.000
	$\sigma = 10$	0	100	0.000

was concave. The estimated translog functions also frequently violated convexity in Scenarios 1 and 3.

6. CONCLUSIONS AND DISCUSSION

CNLS draws its power from the shape constraints that coincide with the standard regularity conditions of the microeconomic theory, avoiding prior assumptions about the functional form or its smoothness. Despite its attractive theoretical properties, applications of CNLS are scarce. In the Introduction, we noted as three major barriers of application: (1) the lack of explicit regression function, (2) computational complexity and (3) difficulty of statistical inference. Our main result is a representation theorem, which shows that the complex, infinite dimensional CNLS problem always has a simple solution characterized by a continuous, piece-wise linear function. Making use of this result, we derived an explicit formulation for a representor function, which can be used as an estimator of the unknown regression function. The representation theorem also enabled us to express the CNLS problem as a QP problem. This facilitates the computation of the CNLS estimators by standard QP algorithms and solver software. Furthermore, a tractable computational procedure enables us to draw statistical inference by applying bootstrap simulations. Thus, we hope that the results of this paper may help to lower the barriers of using CNLS in empirical economic applications. From a methodological point of view, we noted a number of interesting links between CNLS and parallel developments in the literature.

The CNLS approach offers a rich framework for further extensions that fall beyond the scope of this paper. We have restricted attention on estimation of monotonic increasing and concave functions, but the method applies to estimation of monotonic decreasing and/or convex functions in a straightforward fashion. One could relax monotonicity to estimate (inverted) U-shaped functions (such as the Kuznets curves). Relaxing concavity, one arrives at the isotonic regression setting (Barlow et al., 1972). One could also postulate convexity or concavity to apply for a specific range of values, to estimate S-shaped production functions. One might also model homogeneity or homotheticity properties, as briefly suggested in Section 4. The practical implementation of these alternative properties deserves further elaboration.

Another research topic is to adapt the general-purpose regression approach presented here to more specific areas in econometrics, for example, time series or panel data analyses. In the field of production frontier estimation, CNLS has a great potential for unifying the field currently dominated by two separate branches: the parametric regression-based stochastic frontier analysis (SFA) and the deterministic nonparametric DEA (see Kuosmanen, 2006, and Kuosmanen and Kortelainen, 2007, for further discussion). Consumer demand analysis is another area where CNLS has potential to bridge the gap between the nonparametric tests (Afriat, 1967, and Varian, 1982, 1985) and the parametric estimation of demand systems (e.g. Deaton, 1986).

We conclude by noting that the generality of the nonparametric approach does have a price: the rates of convergence are low when the model involves many explanatory variables, which means that large numbers of observations are required to get meaningful estimates. Our Monte Carlo simulations suggest that the method works well when there are relatively few explanatory variables relative to the sample size. In applications with many explanatory variables, the 'curse of dimensionality' could be alleviated by imposing some further semi-parametric structure (e.g. partial linear model). While the asymptotic properties of the nonparametric least squares are well understood, further work is needed to shed light on the impact of the monotonicity, concavity and other inequality constraints on the small sample performance of nonparametic least-squares estimators.

ACKNOWLEDGEMENTS

This paper has benefited from insightful comments by two anonymous reviewers of this Journal. Earlier versions of this paper have been presented at the 29th Annual Meeting of the Finnish Society for Economic Research, Lappeenranta, Finland 1–2 February 2007; the GREMAQ Econometrics Seminar, Toulouse, France, 26 February 2007; 4th Nordic Econometric Meeting, Tartu, Estonia, 24–26 May 2007 and the EEA/ESEM 2007, Budapest, Hungary, 27–31 August, 2007. The author would like to thank Mika Kortelainen, Timo Sipiläinen, Heikki Hella, Christian Bontemps, Pierre Dubois, Thierry Magnac, Martin Browning, Dennis Kristensen, Mogens Fosgerau and other participants to these seminars for useful comments and suggestions. Financial support from the Yrjö Jahnsson Foundation for this research is gratefully acknowledged.

REFERENCES

Afriat, S. N. (1967). The construction of a utility function from expenditure data. *International Economic Review 8*, 67–77.

Banker, R. D. and A. Maindiratta (1992). Maximum likelihood estimation of monotone and concave production frontiers. *Journal of Productivity Analysis 3*, 401–15.

Barlow, R. E., J. M. Bartholomew, J. M. Bremner and H. D. Brunk (1972). *Statistical Inference under Order Restrictions*. New York: John Wiley & Sons.

Deaton, A. (1986). Demand analysis. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics, Volume 3*, 1767–839. Amsterdam: North Holland.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, 1–26.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. *CBMS-NSF Regional Conference Series in Applied Mathematics 38*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficients models. *Annals of Statistics 27*, 1491–518.

Fraser, D. A. S. and H. Massam (1989). A mixed primal-dual bases algorithm for regression under inequality constraints: Application to concave regression. *Scandinavian Journal of Statistics 16*, 65–74.

Goldman, S. M. and P. A. Ruud (1995). Nonparametric multivariate regression subject to constraint. Working paper, University of California at Berkeley.

Greene, W. H. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics 126*, 269–303.

Groeneboom, P., G. Jongbloed and J. A. Wellner (2001). Estimation of convex functions: Characterizations and asymptotic theory. *Annals of Statistics 26*, 1653–98.

Hanson, D. L. and G. Pledger (1976). Consistency in concave regression. *Annals of Statistics 4*, 1038–50.

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

Kuosmanen, T. (2006). Stochastic nonparametric envelopment of data: Combining virtues of SFA and DEA in a unified framework. MTT Discussion Paper 3/2006, Helsinki.

Kuosmanen, T. and M. Kortelainen (2007). Stochastic nonparametric envelopment of data: Cross-sectional frontier estimation subject to shape constraints. Economics Discussion Paper #46, University of Joensuu.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics 19*, 741–59.

Mammen, E. and C. Thomas-Agnan (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics 26*, 239–52.

Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica 59*, 1315–27.

Matzkin, R. L. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*, 239–70.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In R. F. Engle and D.L McFadden (Eds.), *Handbook of Econometrics, Volume 4*, 2523–58. Amsterdam: Elsevier.

Meyer, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference 81*, 13–31.

Meyer, M. C. (2003). A test for linear vs. convex regression function using shape-restricted regression. *Biometrika 90*, 223–32.

Meyer, M. C. (2006). Consistency and power in tests with shape-restricted alternatives. *Journal of Statistical Planning and Inference 136*, 3931–47.

Nemirovskii, A. S., B. T. Polyak and A. B. Tsybakov (1985). Rates of convergence of nonparametric estimates of maximum likelihood type. *Problems of Information Transmission 21*, 258–71.

Rockafellar, R. T. (1970). *Convex Analysis*. NJ: Princeton University Press.

Ruud, P. A. (1996). Restricted least squares subject to monotonicity and concavity constraints, Working paper. University of California at Berkeley.

Sasabuchi, S., M. Inutsuka and D. D. S. Kulatunga (1983). A multivariate version of isotonic regression. *Biometrica 70*, 465–72.

Varian, H. (1982). The nonparametric approach to demand analysis. *Econometrica 50*, 945–73.

Varian, H. (1984). The nonparametric approach to production analysis. *Econometrica 52*, 579–97.

Varian, H. (1985). Nonparametric analysis of optimizing behavior with measurement error. *Journal of Econometrics 30*, 445–58.

Wahba, G. (1990). Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics 59*. Philadelphia: Society for Industrial and Applied Mathematics.

Wu, C. F. (1982). Some algorithms for concave and isotonic regression. *TIMS Studies in Management Science 19*, 105–16.

Yatchew, A. J. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature 36*, 669–721.

Yatchew, A. J. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge: Cambridge University Press.

Yatchew, A. J. and L. Bos (1997). Nonparametric regression and testing in economic models. *Journal of Quantitative Economics 13*, 81–131.

Yatchew, A. J. and W. Härdle (2003). Nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics*, *133*, 579–599.

APPENDIX A: PROOFS

Proof of Theorem 2.1: It is easy to verify that functions $g \in G_2$ satisfy continuity, monotonicity and concavity. Hence $G_2 \subset F_2$. This implies that problem (2.7) involves more stringent constraints than (2.3), and thus we must have $s_g^2 \geq s_f^2$ for any arbitrary data.

Suppose $s_g^2 > s_f^2$. Thus, there exists function $\hat{f} = \arg\min s_f^2$ such that $\hat{f} \in F_2 \backslash G_2$. Define the subdifferential of $\hat{f}$ at point $\mathbf{x}_i \in \mathbb{R}^m$ as

$$\partial \hat{f}(\mathbf{x}_i) = \left\{ \nabla \hat{f}(\mathbf{x}_i) \in \mathbb{R}^m \,\middle|\, \nabla \hat{f}(\mathbf{x}_i) \cdot (\mathbf{x} - \mathbf{x}_i) \leq \hat{f}(\mathbf{x}) - \hat{f}(\mathbf{x}_i) \quad \forall \mathbf{x} \in \mathbb{R}^m \right\}, \tag{A.1}$$

where vector $\nabla \hat{f}(\mathbf{x}_i)$ is referred to as subgradient. Since $\hat{f}$ is continuous, for every observed point $\mathbf{x}_i, i = 1, \ldots, n$ there exists a set of tangent hyperplanes

$$H_i = \left\{ h_i : \mathbb{R}^m \to \mathbb{R} \,\middle|\, h_i(\mathbf{x}) = \hat{f}(\mathbf{x}_i) + \nabla \hat{f}(\mathbf{x}_i) \cdot (\mathbf{x} - \mathbf{x}_i); \nabla \hat{f}(\mathbf{x}_i) \in \partial \hat{f}(\mathbf{x}_i) \right\}. \tag{A.2}$$

Monotonicity implies that

$$\nabla \hat{f}(\mathbf{x}_i) \geq \mathbf{0} \qquad \forall \nabla \hat{f}(\mathbf{x}_i) \in \partial \hat{f}(\mathbf{x}_i), i = 1, \ldots, n. \tag{A.3}$$

Concavity implies that

$$h_i(\mathbf{x}_i) \leq h_k(\mathbf{x}_i) \quad \forall h_i \in H_i; h_k \in H_k; k, i = 1, \ldots, n. \tag{A.4}$$

Note that the objective function of (2.7) depends on the value of $\hat{f}$ at a finite set of points $\mathbf{x}_i$, $i = 1, \ldots, n$. Since $h_i(\mathbf{x}_i) = \hat{f}(\mathbf{x}_i) \quad \forall h_i \in H_i; i = 1, \ldots, n$, without loss of generality, we can represent function $\hat{f}$ by its tangent hyperplanes at these points. Since by assumption $s_g^2 > s_f^2$, there exists at least one tangent hyperplane $h_i \in H_i$ for some $i \in \{1, \ldots, n\}$ that is not feasible for functions $g \in G_2$.

To see that the last claim implies a contradiction, we note that for any given $\nabla \hat{f}(\mathbf{x}_i) \in \partial \hat{f}(\mathbf{x}_i)$, it is possible to set

$$\boldsymbol{\beta}_i = \nabla \hat{f}(\mathbf{x}_i) \tag{A.5}$$

and

$$\alpha_i = h_i(\mathbf{0}). \tag{A.6}$$

With this parametrization, we immediately see that the monotonicity condition (A.3) is equivalent to (2.5). Moreover, the concavity condition (A.4) can be re-written as

$$\hat{f}(\mathbf{x}_i) + \nabla \hat{f}(\mathbf{x}_i) \cdot (\mathbf{x}_i - \mathbf{x}_i) \leq \hat{f}(\mathbf{x}_k) + \nabla \hat{f}(\mathbf{x}_k) \cdot (\mathbf{x}_i - \mathbf{x}_k) \qquad \forall k, i = 1, \ldots, n \tag{A.7}$$

$$\begin{aligned} \Leftrightarrow\ & (\hat{f}(\mathbf{x}_i) - \nabla \hat{f}(\mathbf{x}_i) \cdot \mathbf{x}_i) + \nabla \hat{f}(\mathbf{x}_i) \cdot \mathbf{x}_i \\ & \leq (\hat{f}(\mathbf{x}_k) - \nabla \hat{f}(\mathbf{x}_k) \cdot \mathbf{x}_k) + \nabla \hat{f}(\mathbf{x}_k) \cdot \mathbf{x}_i \qquad \forall k, i = 1, \ldots, n \end{aligned} \tag{A.8}$$

$$\Leftrightarrow \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leq \alpha_k + \boldsymbol{\beta}_k' \mathbf{x}_i \qquad \forall k, i = 1, \ldots, n. \tag{A.9}$$

Inequalities (A.9) are equivalent to the concavity constraints (2.6). Thus, any set of supporting hyperplanes available for functions $\hat{f} \in F_2$ is also available for functions $g \in G_2$. Since the assumption $s_g^2 > s_f^2$ results as a contradiction, we must have $s_g^2 = s_f^2$. □

Proof of Corollary 4.1: Since $G_2 \subset F_2$, then $\hat{g} \in G_2 \Rightarrow \hat{g} \in F_2$. Since problems (2.3) and (3.1) depend on the value of functions f and $\hat{g}$ at a finite set of points $\mathbf{x}_i$, $i = 1, \ldots, n$, Theorem 2.1 directly implies that $\hat{g} \in G_2^* \Rightarrow \hat{g} \in F_2^*$. □

Proof of Theorem 4.1: We start from the lower bound. Note first that $\hat{g}_{\min}$ is continuous, monotonic increasing and concave: $\hat{g}_{\min} \in F_2$. Secondly, note that $\hat{g}_{\min}(\mathbf{x}_i) = \hat{y}_i \quad \forall i \in \{1, \ldots, n\}$. These two observations imply that $\hat{g}_{\min} \in F_2^*$.

Consider an arbitrary $\hat{f} \in F_2^*$, and let $\nabla \hat{f}(\mathbf{x}) \in \partial \hat{f}(\mathbf{x})$ be a subgradient of $\hat{f}$ at an arbitrary point $\mathbf{x} \in \mathbb{R}^m$. The supporting hyperplane theorem (e.g. Rockafellar, 1970) implies that, for all fitted points $(\mathbf{x}_i, \hat{y}_i)$, $i = 1, \ldots, n$, the tangent hyperplanes of a concave $\hat{f}$ at point $\mathbf{x}$ must satisfy

$$\hat{f}(\mathbf{x}) + \nabla \hat{f}(\mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}) \geq \hat{y}_i \quad \forall \nabla \hat{f}(\mathbf{x}) \in \partial \hat{f}(\mathbf{x}). \tag{A.10}$$

Using $\alpha = \hat{f}(\mathbf{x}) - \nabla \hat{f}(\mathbf{x}) \cdot \mathbf{x}$ and $\boldsymbol{\beta} = \nabla \hat{f}(\mathbf{x})$, function $\hat{g}_{\min}$ can be re-written as

$$\hat{g}_{\min}(\mathbf{x}) = \min_{\substack{\alpha \in \mathbb{R} \\ \boldsymbol{\beta} \in \mathbb{R}^m}} \left\{ \alpha + \boldsymbol{\beta}' x | \alpha + \boldsymbol{\beta}' x_i \geq \hat{y}_i \quad \forall i = 1, \ldots, n \right\} \tag{A.11}$$

$$= \min_{\hat{f}} \left\{ (\hat{f}(\mathbf{x}) - \nabla \hat{f}(\mathbf{x}) \cdot \mathbf{x}) + \nabla \hat{f}(\mathbf{x}) \cdot \mathbf{x} \,\middle|\, (\hat{f}(\mathbf{x}) - \nabla \hat{f}(\mathbf{x}) \cdot \mathbf{x}) + \nabla \hat{f}(\mathbf{x}) \cdot \mathbf{x}_i \geq \hat{y}_i \, \forall i = 1, \ldots, n \right\} \tag{A.12}$$

$$= \min_{\hat{f}} \left\{ \hat{f}(\mathbf{x}) \,\middle|\, \hat{f}(\mathbf{x}) + \nabla \hat{f}(\mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}) \geq \hat{y}_i \quad \forall i = 1, \ldots, n \right\}. \tag{A.13}$$

Therefore, $\hat{g}_{\min}(\mathbf{x}) = \min_f \{ f(\mathbf{x}) | f \in F_2^* \}$. This completes the first part of the proof.

Consider next the upper bound

$$\hat{g}_{\max}(\mathbf{x}) = \max_{\phi \in \mathbb{R}, \boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^{m \times n}} \left\{ \phi \,\middle|\, \phi \leq \alpha_i + \boldsymbol{\beta}_i' \mathbf{x} \, \forall i; \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i = \hat{y}_i \, \forall i; \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_h \geq \hat{y}_h \quad \forall h \neq i \right\}. \tag{A.14}$$

Note that there must exist some observation $i = \arg\min_{i\in\{1,\ldots,n\}} \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}$ for which the constraint $\phi \leq \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}$ holds as equality. Therefore, the upper bound (A.14) can be equivalently written as

$$\hat{g}_{\max}(\mathbf{x}) = \min_{i\in\{1,\ldots,n\}} \left(\max_{\alpha_i\in\mathbb{R},\boldsymbol{\beta}_i\in\mathbb{R}^m} \left\{ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x} \,\middle|\, \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i = \hat{y}_i; \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_h \geq \hat{y}_h \quad \forall h \neq i \right\} \right). \tag{A.15}$$

This minimax formulation reveals that function $\hat{g}_{\max}$ is not concave, and thus, $\hat{g}_{\max} \notin F_2^*$. Using again $\alpha = \hat{f}(\mathbf{x}) - \nabla\hat{f}(\mathbf{x})\cdot\mathbf{x}$ and $\boldsymbol{\beta} = \nabla\hat{f}(\mathbf{x})$, function $\hat{g}_{\max}$ can be expressed as

$$\hat{g}_{\max}(\mathbf{x}) = \min_{i\in\{1,\ldots,n\}} \left(\max_{\hat{f}_i} \left\{ \hat{f}_i(\mathbf{x}) \,\middle|\, \hat{f}_i(\mathbf{x}_i) = \hat{y}_i; \hat{f}_i(\mathbf{x}) + \nabla\hat{f}_i(\mathbf{x})\cdot(\mathbf{x}_h - \mathbf{x}) \geq \hat{y}_h \quad \forall h \neq i \right\} \right). \tag{A.16}$$

Consider first the embedded maximization problem in (A.16). The problem is analogous to (A.13), except that we have replaced the minimization by maximization, and we force the tangent line to pass through a given point $(\mathbf{x}_i, \hat{y}_i)$. Constraint $\hat{f}_i(\mathbf{x}_i) = \hat{y}_i$ is necessary in (A.16), because otherwise the problem would be unbounded. In essence, the embedded maximization problem of (A.16) finds the tangent hyperplane through a fixed point $(\mathbf{x}_i, \hat{y}_i)$ with the largest value at point $\mathbf{x}$. The first-order conditions of problem $\max_f \left\{ f(\mathbf{x}) \,\middle|\, f \in F_2^* \right\}$ imply that the optimum is achieved at one of the points $\hat{f}_i(\mathbf{x})$ on the tangent hyperplane of some $i \in \{1,\ldots,n\}$.

To see why the optimum must be the minimum value over $i \in \{1,\ldots,n\}$, consider observations $i, j \in \{1,\ldots,n\}$ and let $\hat{f}_j^*(\mathbf{x})$, $\hat{f}_i^*(\mathbf{x})$ be the optimal solutions to the embedded maximization problem of (A.16) such that $\hat{f}_j^*(\mathbf{x}) > \hat{f}_i^*(\mathbf{x})$. The maximizing property of $\hat{f}_i^*(\mathbf{x})$ implies that there exists a subset $S \subset \{1,\ldots,n\}$ such that $\hat{f}_i^*(\mathbf{x}) + \nabla\hat{f}_i^*(\mathbf{x})\cdot(\mathbf{x}_s - \mathbf{x}) = \hat{y}_s \quad \forall s \in S$ (i.e. the constraints of the maximization problem in (A.16) are binding for all observations $s \in S$). But then it is possible to construct point $(\vec{\mathbf{x}}, \vec{y})$ as a convex combination $\vec{\mathbf{x}} = \sum_{s\in S}\lambda_s\mathbf{x}_s + \lambda_j\mathbf{x}$, $\vec{y} = \sum_{s\in S}\lambda_s\hat{y}_s + \lambda_j\hat{f}_j(\mathbf{x})$, $\sum_{s\in S}\lambda_s + \lambda_j = 1, \lambda_s\lambda_j \geq 0 \quad \forall s \in S$ such that $\vec{\mathbf{x}} = \mathbf{x}_i$ and $\vec{y} > \hat{y}_i$. Therefore, choosing point $(\mathbf{x}, \hat{f}_j^*(\mathbf{x}))$ violates of the concavity postulate. As this argument applies to any observations $i, j \in \{1,\ldots,n\}$, the only feasible solution is obtained by minimizing over observations $i \in \{1,\ldots,n\}$. Therefore, $\hat{g}_{\max}(\mathbf{x}) = \max_f\{f(\mathbf{x}) | f \in F_2^*.\}$. □

REMARK A.1. The dual problem of (4.4) provides some further intuition especially to the first part of the proof of Theorem 4.1. The dual formulation can be written as

$$\hat{g}_{\min}(\mathbf{x}) = \max_{\mathbf{z}\in\mathbb{R}_+^n} \left\{ \sum_{i=1}^n z_i\hat{y}_i \,\middle|\, \mathbf{x} \geq \sum_{i=1}^n z_i\mathbf{x}_i; \sum_{i=1}^n z_i = 1 \right\}, \tag{A.17}$$

where z_i represents the weight assigned to observation i. In essence, the dual problem (A.17) maximizes the weighted average of the fitted values $\hat{y}_i$, subject to the constraint that the weighted average of the observed explanatory variables is less than or equal to the given level of $\boldsymbol{x}$. Below the observed range, problem (A.17) is infeasible and is hence assigned the value $-\infty$, whereas above the observed range $\hat{g}_{\min}(\mathbf{x}) = \max_n\{\hat{y}_n\}$.

APPENDIX B: GAMS CODE FOR THE CNLS REGRESSION (100 OBSERVATIONS, TWO EXPLANATORY VARIABLES)

```
SETS
k              observations /1*100/
alias(k,m);

PARAMETERS
Y(k)           dependent variable
X1(k)          value of explanatory variable 1 in observation k
X2(k)          value of explanatory variable 2 in observation k

VARIABLES
E(k)           residual of k
A(k)           constant
SS             sum of square of errors;

POSITIVE VARIABLES
B1(k)          beta 1 coefficients
B2(k)          beta 2 coefficients;

EQUATIONS
QSSE                    objective = sum of squares of residuals
QREGRESSION(k)          regression equation
QCONCAVITY(k,m)         concavity constraint;
QSSE..                  SS = e = sum(k,E(k)*E(k));
QREGRESSION(k)..        Y(k) = e = A(k)+B1(k)*X1(k)+B2(k)*X2(k)+E(k);
QCONCAVITY(k,m)..       Y(m) = l = A(k)+B1(k)*X1(m)+B2(k)*X2(m)+E(m);
MODEL CNLS /all/
SOLVE CNLS using NLP Minimizing SS;
```