OPERATIONS RESEARCH
Vol. 58, No. 1, January–February 2010, pp. 149–160
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5801 | 0149

informs

DOI 10.1287/opre.1090.0722

Data Envelopment Analysis as Nonparametric Least-Squares Regression

Timo Kuosmanen
Economic Research Unit, MTT Agrifood Research Finland, 00410 Helsinki, Finland, and
Department of Business Technology, Helsinki School of Economics, 00101 Helsinki, Finland,
timo.kuosmanen@hse.fi

Andrew L. Johnson
Department of Industrial and Systems Engineering, Texas A&M University,
College Station, Texas 77843, ajohnson@tamu.edu

Data envelopment analysis (DEA) is known as a nonparametric mathematical programming approach to productive efficiency analysis. In this paper, we show that DEA can be alternatively interpreted as nonparametric least-squares regression subject to shape constraints on the frontier and sign constraints on residuals. This reinterpretation reveals the classic parametric programming model by Aigner and Chu [Aigner, D., S. Chu. 1968. On estimating the industry production function. *Amer. Econom. Rev.* **58** 826–839] as a constrained special case of DEA. Applying these insights, we develop a nonparametric variant of the corrected ordinary least-squares (COLS) method. We show that this new method, referred to as corrected concave nonparametric least squares (C^2NLS), is consistent and asymptotically unbiased. The linkages established in this paper contribute to further integration of the econometric and axiomatic approaches to efficiency analysis.

Subject classifications: frontier estimation; mathematical programming; nonparametric estimation; performance measurement; benchmarking.
Area of review: Decision Analysis.
History: Received May 2008; revisions received May 2008, November 2008; accepted April 2009. Published online in *Articles in Advance* October 7, 2009.

1. Introduction

Data envelopment analysis (DEA) is an axiomatic, mathematical programming approach to productive efficiency analysis of firms and other decision-making units. Originating from the work of Farrell (1957), DEA's current popularity is largely due to the seminal paper by Charnes et al. (1978). Thousands of DEA studies have been reported in application areas including agriculture, education, financial institutions, health care, public sector firms, etc. DEA's real-world relevance, diffusion, and global acceptance are evident from literature studies such as Seiford (1996) and Gattoufi et al. (2004).

DEA's chief advantage compared to econometric, regression-based tools is its nonparametric treatment of the frontier. Relying on general axioms of production theory, e.g., monotonicity, convexity, and homogeneity, DEA does not assume any particular functional form. Its direct, data-driven approach is essential for communicating the results of efficiency analysis to decision makers. However, DEA is often criticized for its deterministic, nonstatistical nature. Schmidt (1985) phrased the criticism as follows:

> "I am very skeptical of non-statistical measurement exercises, certainly as they are now carried out and perhaps in any way in which they could be carried out.... I see no virtue whatever in a non-statistical approach to data." p. 296

Banker (1993) was among the first to respond to Schmidt's critique by identifying conditions under which DEA estimators are statistically consistent and have a maximum likelihood (ML) interpretation. At present, the formal statistical foundation of DEA estimators is well established, including the asymptotic theory and rates of convergence as well as methods for statistical inference (see, e.g., Simar and Wilson 2008 for comprehensive surveys of this work). In addition, extensions to DEA have been proposed to improve its robustness to data errors and outliers (e.g., stochastic DEA, DEA+, chance-constrained DEA, and robust DEA frontiers, developed by Banker et al. 1991, Gstach 1998, Cooper et al. 1996, and Daraio and Simar 2007, respectively). However, these approaches are still nonstatistical in the sense of Schmidt (quoted above), and they do not allow for a genuine probabilistic treatment of stochastic noise in the observed data. Most importantly, a large conceptual and philosophical gap remains between the mathematical-programming-based DEA and the regression-based econometric approaches (cf., e.g., Cooper et al. 2004).

This paper contributes to bridging the conceptual gap by demonstrating that DEA can be recast as least-squares regression. More specifically, we show that the standard (output-oriented, variable returns to scale) DEA model can be formulated as nonparametric least-squares regression

subject to shape constraints (monotonicity and concavity) on the frontier and a sign constraint regarding the regression residuals. Whereas Banker (1993) has earlier shown that DEA has an ML interpretation (cf. also Banker and Maindiratta 1992), the least-squares interpretation established in this paper is a new result that further enhances the statistical foundation of DEA. It is worth emphasizing that ML estimation generally requires the inefficiency terms to be identically and independently distributed according to some specific probability density function, whereas least-squares estimation does not require such assumptions. Moreover, although the ML interpretation of DEA applies to a broad class of inefficiency distributions, its practical usefulness is limited due to the classic incidental parameters problem. As a result, the usual asymptotic properties of ML estimators do not apply to DEA estimators (see Schmidt 1985 and Banker 1993). By contrast, the direct link between DEA and least-squares regression established in this paper can be utilized in many ways in the nonparametric statistical estimation of the axiomatic production model, as will be demonstrated below. Further, this result sheds new light on the connections between alternative frontier estimation techniques. Several connections arise via this result; particularly, this result implies that DEA can be obtained as a nonparametric generalization of the classic parametric programming model by Aigner and Chu (1968).

Applying these insights, we develop a new nonparametric variant of the correct ordinary least squares (COLS) approach (Greene 1980). We call this new method *Corrected concave nonparametric least squares* (C^2NLS). If the data-generating process is deterministic and the inefficiency terms are identically and independently distributed, our C^2NLS method offers certain advantages to the traditional DEA. Whereas DEA spans the efficient frontier on a few influential data points, C^2NLS uses the information in all observations for estimating the frontier. Thus, it is less vulnerable to small-sample error. We show that the estimates from C^2NLS are consistent and asymptotically unbiased, and yield smaller bias and mean-squared error than the corresponding DEA efficiency estimators as the dimensions of the input-output space increase beyond the single-input, single-output case.

Although the C^2NLS method is not designed for settings where data are perturbed by stochastic noise, the evidence from Monte Carlo simulations suggests that the efficiency rankings obtained with C^2NLS are generally more robust to noise than the DEA rankings are, especially when the variance of the noise component is relatively small compared to the variance of the inefficiency component. In this respect, we see the development of the least-squares interpretation for DEA as a pivotal first step towards integrating truly stochastic inefficiency and noise terms to the nonparametric approach to productive efficiency analysis. This line of research is pursued further in the follow-up paper by Kuosmanen and Kortelainen (2007), which departs from Step 1 of the C^2NLS method to estimate the conditional expected value of the inefficiency term in a probabilistic fashion based on the distribution of nonparametric least-squares residuals.

The remainder of this paper is organized as follows: §2 introduces the necessary notation and describes the CNLS estimation method. In §3, the connection between DEA and least-squares regression is established. Utilizing these insights, the corrected nonparametric least-squares method is introduced and discussed in §4. Section 5 presents some evidence from Monte Carlo simulations. Section 6 gives some concluding remarks. An exact description of the PP and COLS models referred to in this paper is presented in Online Appendix 1. Formal proofs of all mathematical theorems are presented in Online Appendix 2. An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

2. Models of Production

2.1. Classification

Consider the standard multiple-input, single-output, cross-sectional model in production economics:

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

where y_i denotes the output of firm i, $f\colon \Re_+^m \to \Re_+$ is the production function that characterizes the production technology, $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})'$ is the input vector of firm i, and ε_i is the disturbance term that represents the deviation of firm i from the frontier. Different models of productive efficiency analysis can be classified according to how one specifies the production function f and disturbances ε_i.

Generally, models are classified as *parametric* or *nonparametric* depending on the specification of the production function f. Parametric models postulate a priori a specific functional form for f (e.g., Cobb-Douglas, translog, etc.) and subsequently estimate its unknown parameters. Nonparametric models assume that f satisfies certain regularity axioms (e.g., monotonicity and concavity), but no particular functional form is assumed.

Models can also be classified as *neoclassical* or *frontier* depending on the interpretation of the disturbance term ε_i (cf., Kuosmanen and Fosgerau 2009). The neoclassical model assumes that all firms are efficient and disturbances ε_i are random, uncorrelated noise terms. Frontier models typically assume that all deviations from the frontier are attributed to inefficiency, which implies that $\varepsilon_i \leqslant 0 \;\; \forall\, i = 1, \ldots, n$. For brevity, we omit the stochastic frontier models (Aigner et al. 1977, Meeusen and Vandenbroeck 1977), where ε_i are interpreted as composite error terms that include both inefficiency and noise components (see, e.g., Kuosmanen 2006 and Kuosmanen and Kortelainen 2007 for further discussion of stochastic, nonparametric frontier models).

Table 1. Classification of production models and the linkages established in this paper.

	Parametric	Nonparametric
Central tendency	OLS Cobb and Douglas (1928)	CNLS Hildreth (1954) Hanson and Pledger (1976)
Frontier; sign constraints	PP Aigner and Chu (1968) Timmer (1971)	DEA Farrell (1957) Charnes et al. (1978)
Frontier; 2-stage estimation	COLS Winsten (1957) Greene (1980)	C^2NLS This paper

Table 1 combines the criteria described above to identify six alternative model variants, together with some canonical references. On the parametric side, OLS refers to *ordinary least squares*, PP means *parametric programming*, and COLS is *corrected ordinary least squares* (see Online Appendix 1 for details regarding PP and COLS). On the nonparametric side, CNLS refers to *convex nonparametric least squares* (§2.2), DEA is *data envelopment analysis* (§3), and C^2NLS is a new approach called *corrected convex nonparametric least squares* (§4). Although the traditional parametric models are well established in productive efficiency analysis, the nonparametric least-squares techniques are less well known in this literature. Thus, a brief introduction to CNLS is provided in the following subsection.

The original developments in this paper can be summarized in terms of Table 1 as follows:

(1) We establish formal links between DEA and CNLS regression, showing that DEA is a sign-constrained special case of CNLS.

(2) As a corollary to this result, we find that DEA is a nonparametric generalization of PP.

(3) We develop a new nonparametric generalization of COLS, referred to as C^2NLS.

2.2. Concave Nonparametric Least Squares (CNLS)

Nonparametric regression techniques that do not require any prior assumption about the functional form of the regression function come in many varieties (see, e.g., Yatchew 1998, 2003 for a comprehensive survey). Nonparametric least squares subject to continuity, monotonicity, and concavity constraints arose from work by Hildreth (1954).[1] Following Kuosmanen (2008), we refer to this approach as concave nonparametric least squares (CNLS). CNLS is based on the assumptions that the regression function f to be estimated belongs to the set of continuous, monotonic increasing, and globally concave functions, denoted henceforth by F_2, and the disturbances $\boldsymbol{\varepsilon} = (\varepsilon_1 \cdots \varepsilon_n)'$ satisfy the Gauss-Markov assumptions (i.e., $E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2\mathbf{I}$, $\sigma < \infty$). Note that in contrast to maximum likelihood estimation, in least-squares estimation $\boldsymbol{\varepsilon}$ need not be identically and independently distributed (i.i.d.); $\boldsymbol{\varepsilon}$ are only assumed to be uncorrelated with inputs $\mathbf{X}$ and with each other. Further, in contrast to the kernel regression and spline-smoothing techniques, CNLS does not require specification of any smoothing or bandwidth parameters.

The CNLS problem is to find $f \in F_2$ that minimizes the sum of squared deviations, formally,

$$\min_{f,\boldsymbol{\varepsilon}} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \;\middle|\; y_i = f(\mathbf{x}_i) + \varepsilon_i \;\forall i = 1,\ldots,n;\; f \in F_2 \right\}. \tag{2}$$

Note that the family F_2 includes an infinite number of functions, which makes problem (2) a challenging infinite-dimensional problem. Earlier single regressor CNLS algorithms (e.g., Fraser and Massam 1989, Meyer 1999) require that the data are sorted in ascending order according to the scalar-valued regressor x. However, such sorting is not possible in the general multiple-regression setting where $\mathbf{x}$ is a vector.

To estimate the CNLS problem (2) in the general multi-input setting, Kuosmanen (2008) has shown that the family F_2 can be equivalently represented by a family of piecewise-linear functions characterized by the celebrated Afriat's Theorem (Afriat 1967, 1972).[2] More specifically, we can model the values of f by using a system of supporting hyperplanes, imposing the concavity constraint by means of Afriat inequalities. Applying these insights, we may rewrite the infinite-dimensional problem (2) as the following finite-dimensional quadratic programming (QP) problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \left\{ \sum_{I=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{array}{ll} y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i & \forall i = 1,\ldots,n; \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \leqslant \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i & \forall h, i = 1,\ldots,n; \\ \boldsymbol{\beta}_i \geqslant \mathbf{0} & \forall i = 1,\ldots,n \end{array} \right\}. \tag{3}$$

The least-squares problems (2) and (3) are equivalent in the sense that the optimal objective values of the two problems are equal for any real-valued data set (Kuosmanen 2008).

In problem (3), the first constraint estimates α_i and $\boldsymbol{\beta}_i$ parameters for each observation; thus, n different regression lines are estimated instead of fitting one regression line to the cloud of observed points, as in OLS. Note that by replacing the constraint $\boldsymbol{\beta}_i \geqslant \mathbf{0}$ $\forall i$ by constraints $\alpha_i = \alpha_j$, $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ $\forall i, j = 1,\ldots,n$, we obtain the standard OLS problem. These n estimated lines can be interpreted as tangent lines to the unknown production function f. The slope coefficients $\boldsymbol{\beta}_i$ represent the marginal products of inputs (i.e., the subgradients $\nabla f(\mathbf{x}_i)$). The second constraint imposes concavity by applying a system of Afriat inequalities; these inequalities are the key to modeling concavity constraints in the general multiple-regressor setting. The third constraint imposes monotonicity.

Given the estimated coefficients $(\alpha_i, \boldsymbol{\beta}_i)$ from (3), the following explicit estimator of f can be constructed:

$$f^{CNLS}(\mathbf{x}) = \min_{i \in \{1,\ldots,n\}} \{\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}\}. \tag{4}$$

In principle, estimator f^{CNLS} consists of n hyperplane segments, but in practice the estimated coefficients $(\alpha_i, \boldsymbol{\beta}_i)$ are clustered to a relatively small number of alternative values: The number of different hyperplane segments is usually much lower than n.[3] It is important to observe that if we denote the set of functions that minimize the original CNLS problem (2) by F_2^*, it becomes easy to show that $f^{CNLS} \in F_2^*$ for any finite real-valued data set (Theorem 3.2 of Kuosmanen 2008).

The essential statistical properties of the CNLS estimators are well understood. The maximum likelihood interpretation of CNLS was 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) showed that CNLS achieves the optimal nonparametric rate of convergence $\mathrm{Op}(n^{-4/(4+m)})$, where n is the number of observations and m is the number of regressors (cf. Stone 1980). Imposing further smoothness assumptions or derivative bounds has been explored by Mammen (1991), Yatchew (1998), and Mammen and Thomas-Agnan (1999). Groeneboom et al. (2001) have derived the asymptotic distribution of a CNLS estimator at a fixed point.

3. DEA as Nonparametric Least Squares

For a production function f estimated under the maintained assumptions of monotonicity and concavity (i.e., the DEA production function), the variable returns to scale (VRS) DEA estimator of f can be formally defined as (Afriat 1972, Banker 1993)[4]

$$f^{DEA}(\mathbf{x}) = \max_{\lambda \in \mathbb{R}^n_+} \left\{ y \;\middle|\; y = \sum_{h=1}^{n} \lambda_h y_h;\ \mathbf{x} \geqslant \sum_{h=1}^{n} \lambda_h \mathbf{x}_h;\ \sum_{h=1}^{n} \lambda_h = 1 \right\}. \quad (5)$$

Multipliers λ_i are referred to as intensity weights (used for constructing convex combinations of the observed firms). Substituting f in (1) by the DEA estimator (5), we see that the DEA efficiency estimate ε_i^{DEA} for firm i is obtained as the optimal solution to the following linear programming (LP) problem:

$$\varepsilon_i^{DEA} = \min_{\lambda,\, \varepsilon} \left\{ \varepsilon \;\middle|\; y_i = \sum_{h=1}^{n} \lambda_h y_h + \varepsilon;\ \mathbf{x}_i \geqslant \sum_{h=1}^{n} \lambda_h \mathbf{x}_h;\ \sum_{h=1}^{n} \lambda_h = 1;\ \lambda_h \geqslant 0\ \forall h = 1, \ldots, n \right\}. \quad (6)$$

Note that the DEA formulation (6) differs from the standard output-oriented VRS DEA model by Banker et al. (1984) in a subtle way. Whereas problem (6) is consistent with the additive single-output specification of (1), Banker et al. measure efficiency in the multiplicative form using the LP problem (here adapted to the single-output setting)

$$\theta_i^{DEA} = \max_{\lambda,\, \theta} \left\{ \theta \;\middle|\; \theta y_i \leqslant \sum_{h=1}^{n} \lambda_h y_h;\ \mathbf{x}_i \geqslant \sum_{h=1}^{n} \lambda_h \mathbf{x}_h;\ \sum_{h=1}^{n} \lambda_h = 1;\ \lambda_h \geqslant 0\ \forall h = 1, \ldots, n \right\}. \quad (7)$$

The LP problem (7) is consistent with the radial Farrell output efficiency measure. Despite the subtle difference in the scale of measurement, formulations (6) and (7) are equivalent in the following sense:

LEMMA 3.1. *In the single-output setting, the additive DEA efficiency measure* (6) *is equivalent to the multiplicative DEA efficiency measure* (7) *in the sense that*

$$\theta_i^{DEA} = 1 - \varepsilon_i^{DEA}/y_i \quad \forall i = 1, \ldots, n. \quad (8)$$

Both problems (6) and (7) measure efficiency relative to the same DEA frontier characterized by (5).

To establish the least-squares interpretation for DEA, it is first useful to recall the PP method developed by Aigner and Chu (1968) (see Online Appendix 1). In essence, the PP problem is the standard OLS problem augmented with the additional sign constraint on the residuals. Although the PP model is deterministic like DEA, it provides legitimate means for statistical estimation: Schmidt (1976) has shown that PP provides maximum likelihood estimators if the inefficiency terms ε are identically and independently distributed with the half-normal probability density (see Online Appendix 1 for details).

As an obvious nonparametric counterpart to PP, we may consider a sign-constrained variant of the CNLS problem, formally,

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\varepsilon}} \left\{ \sum_{I=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{array}{l} \varepsilon_i \leqslant 0 \quad \forall i = 1, \ldots, n; \\ y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i \quad \forall i = 1, \ldots, n; \\ \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leqslant \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i \quad \forall h, i = 1, \ldots, n; \\ \boldsymbol{\beta}_i \geqslant \mathbf{0} \quad \forall i = 1, \ldots, n \end{array} \right\}. \quad (9)$$

Comparing problems (3) and (9), we see that (9) is simply the sign-constrained variant of the CNLS problem. The monotonicity and concavity assumptions on frontier f still hold, but the Gauss-Markov assumptions on $\boldsymbol{\varepsilon}$ are violated due to the sign constraint in (9). In model (9) we do not require that $\boldsymbol{\varepsilon}$ are uncorrelated (we return to this point in more detail in §4). On the other hand, we note that the QP problem (9) differs from the classic PP model in that it does not assume any particular functional form, but rather builds upon the axioms of monotonicity and concavity, similar to CNLS and DEA. In this sense, model (9) is a nonparametric generalization of PP. Interestingly, this hybrid model of PP and CNLS is equivalent to the standard DEA model.

THEOREM 3.1. *For all real-valued data, the sign-constrained nonparametric least-squares problem* (9) *is equivalent to the DEA model* (6) *in the sense that* $\varepsilon_i^{DEA} = \varepsilon_i^*$ *for all* $i = 1, \ldots, n$, *where* ε_i^*, $i = 1, \ldots, n$, *are obtained as the optimal solution to problem* (9). *Both problems* (6) *and* (9) *measure efficiency relative to the same DEA frontier characterized by* (5).

This result is important for several reasons. From a methodological perspective, Theorem 3.1 elaborates and further enhances the statistical foundation of DEA. As a response to Schmidt's (1985) critique, quoted in the introduction, Banker (1993) has shown that ε_i^{DEA} are maximum likelihood estimators for a broad class of inefficiency distributions, including the exponential and half-normal distributions considered by Schmidt (1976, 1985). Interestingly, whereas Banker's results show that DEA has a statistical justification analogous to that of PP, our Theorem 3.1 implies that PP is in fact a constrained special case of DEA. The connection between DEA and PP is evident from comparing formulation (9) with the PP formulation (A1.2) in Online Appendix 1.

In our interpretation, Theorem 3.1 demonstrates that the conceptual and philosophical barriers between DEA and regression-based approaches are considerably lower than what has been assumed before. Although DEA and regression analysis have been thought to be very different and incompatible (see, e.g., Cooper et al. 2004, or Schmidt 1985, quoted in the introduction), it is possible to approach DEA as a sign-constrained variant of nonparametric least-squares regression. We hope that the established linkages could contribute to the further integration of the parametric and nonparametric approaches towards a unified framework of productive efficiency analysis.

From a practical point of view, the development of this connection opens up new avenues for integrating tools from econometrics and the axiomatic, mathematical-programming-based approaches. For example, DEA currently lacks a meaningful goodness-of-fit statistic. Given the least-squares formulation derived in this paper, we could apply the coefficient of determination from regression analysis, specifically,

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (\varepsilon_i^{DEA})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= 1 - \frac{\sum_{i=1}^n ((1-\theta_i) y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{10}$$

for measuring the proportion of output variation that is explained by the DEA frontier. Although this variance decomposition can be applied to any regression model, including DEA, we must stress that, in contrast to OLS, DEA does not maximize R^2. Consequently, negative R^2 values are possible for DEA estimators.

Some other promising developments expected from the integration of econometric and the axiomatic, mathematical programming approaches include: (1) decomposition of the error term in nonparametric models to distinguish inefficiency from random noise in a probabilistic manner (cf. Kuosmanen and Kortelainen 2007), (2) incorporation of variables exogenous to a firm in a single-stage nonparametric model, (3) incorporation of multiple-output extensions such as the stochastic distance function to multioutput DEA, and (4) incorporation of a method for statistical inference (hypotheses testing, confidence intervals) available in the regression literature into the nonparametric efficiency analysis. Although these developments fall beyond the scope of the present paper, none of these potentially path-breaking improvements are possible without the connection established in Theorem 3.1.

From a computational point of view, two points are worth noting. First, observe that the DEA efficiency estimates are obtained as the optimal ε_i from (9), not the squared values ε_i^2. Interestingly, replacing the squared terms ε_i^2 by ε_i in the objective function of (9) will not change the optimal value; the least-squares criterion is equivalent to the least absolute deviation (LAD) criterion in this sign-constrained case. This is not generally true if we relax the sign constraint on ε_i as in (3). Second, the QP formulation (9) computes the DEA efficiency measures for all firms simultaneously, whereas in DEA models (6) and (7) the efficiency measures are solved separately for each firm in the sample. In the conventional DEA setting, solving n small, independent LP problems is typically faster and more economical that solving a single large LP problem.[5] However, saving computation time is not our objective. Importantly, we need to include all n firms into a single large problem if we want to model interdependencies in the efficiency estimates across firms (compare with Kuosmanen et al. 2006). For example, note that the QP problem (3) that excludes the sign constraint on residuals cannot generally be broken down to independent subproblems.

Regarding generality of the result, we observe that Theorem 3.1 is by no means restricted to the VRS technology. By augmenting problem (9) with additional linear constraints on the intercept terms α_i, we can derive analogous results for the other standard specifications of returns to scale, including constant $[\alpha_i = 0 \ \forall i = 1, \ldots, n]$, nonincreasing $[\alpha_i \geqslant 0 \ \forall i = 1, \ldots, n]$, or nondecreasing $[\alpha_i \leqslant 0 \ \forall i = 1, \ldots, n]$ returns to scale. The result can also be extended to the input-oriented efficiency measure and the multioutput technology by rephrasing the model in the single-input setting with an additive inefficiency term.

In conclusion, we have demonstrated that DEA can be expressed as a sign-constrained CNLS problem, which generalizes the PP model by relaxing the parametric assumption. Applying insights from Theorem 3.1, we may develop nonparametric generalizations to other models established in the parametric literature. The next section develops a nonparametric generalization of the corrected ordinary least squares (COLS) model.

4. Corrected Concave Nonparametric Least Squares (C^{2}NLS)

Corrected nonparametric least squares (C^{2}NLS), to be developed next, is a new nonparametric variant of the COLS model in which nonparametric least squares subject to monotonicity and concavity constraints replace the first-stage parametric OLS regression. The C^{2}NLS model assumes that the regression f is monotonic increasing and

globally concave, the inefficiencies $\boldsymbol{\varepsilon}$ are identically and independently distributed (i.i.d.) with mean μ and a finite variance σ^2, and that the inefficiencies $\boldsymbol{\varepsilon}$ are uncorrelated with inputs $\mathbf{X}$. Note that the i.i.d. assumption on $\boldsymbol{\varepsilon}$ implies the following modified set of Gauss-Markov conditions: $E(\boldsymbol{\varepsilon} - \mu\mathbf{1} \mid \mathbf{X}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2\mathbf{I}$. These conditions can be exploited for constructing an efficient least-squares estimator.

Like COLS, the C^{2}NLS method is implemented in two stages, which can be stated as follows:

Stage 1: Estimate $E(y_i \mid \mathbf{x}_i)$ by solving the CNLS problem (3). Denote the CNLS residuals by ε_i^{CNLS}.

Stage 2: Shift the residuals analogous to the COLS procedure; the C^{2}NLS efficiency estimator is

$$\hat{\varepsilon}_i^{C2NLS} = \varepsilon_i^{CNLS} - \max_h \varepsilon_h^{CNLS}, \tag{11}$$

where values of $\hat{\varepsilon}_i^{C2NLS}$range from $[0, -\infty]$ with 0 indicating efficient performance. Similarly, we adjust the CNLS intercepts α_i as

$$\hat{\alpha}_i^{C2NLS} = \alpha_i^{CNLS} + \max_h \varepsilon_h^{CNLS}, \tag{12}$$

where α_i^{CNLS} is the optimal intercept for firm i in (3) and $\hat{\alpha}_i^{C2NLS}$ is the C^{2}NLS estimator. Slope coefficients $\hat{\boldsymbol{\beta}}_i$ for C^{2}NLS are obtained directly as the optimal solution to (3).

The key difference between COLS and C^{2}NLS concerns Stage 1, where COLS uses parametric OLS, whereas C^{2}NLS uses nonparametric CNLS. In this respect, COLS can be seen as a restricted special case of C^{2}NLS.

Concerning the statistical properties of C^{2}NLS, we can apply the results by Hanson and Pledger (1976) and Greene (1980) to prove the following asymptotic result.

THEOREM 4.1. *For any sequence of independent observations* $\mathbf{X}$, $\mathbf{y}$ *generated by production function* $f \in F_2$ *and identically and independently distributed inefficiency terms* $\varepsilon_i \leqslant 0$ *that are uncorrelated with* $\mathbf{X}$ *and have a positive density at* $\varepsilon_i = 0$, *the C^{2}NLS efficiency estimator is statistically consistent. Specifically,*

$$\lim_{n\to\infty} \hat{\varepsilon}_i^{C2NLS} = \varepsilon_i \quad \forall\, i = 1, \ldots, n. \tag{13}$$

Consistency is a generally desirable property for any efficiency estimator. Consistency of DEA has been proven by Banker (1993) and Korostelev et al. (1995), and consistency of C^{2}NLS is established in Theorem 4.1 above. The i.i.d. assumption on inefficiency terms $\boldsymbol{\varepsilon}$ made in Theorem 4.1 is worth elaborating. Note that consistency of DEA estimators (including the least-squares formulation (9)) requires that the data set is a random sample of n independent observations, but the assumption of i.i.d. inefficiency terms is not required; DEA allows $\boldsymbol{\varepsilon}$ to be correlated with inputs and/or with each other. However, the ML interpretation of DEA (Banker 1993) does require the same i.i.d. assumption as in Theorem 4.1. On the other hand, recall that the CNLS estimator used in Stage 1 only requires that the inefficiency terms $\boldsymbol{\varepsilon}$ are uncorrelated with the inputs and with each other. The somewhat stronger i.i.d. assumption made in Theorem 4.1 is necessary for shifting the frontier in Stage 2.

It is illustrative to briefly consider how possible violations of the assumptions on $\boldsymbol{\varepsilon}$ stated in Theorem 4.1 would influence the performance of C^{2}NLS. First, if $E(\boldsymbol{\varepsilon} - \mu\mathbf{1} \mid \mathbf{X}) \neq \mathbf{0}$, then the CNLS estimator used in Stage 1 is biased and inconsistent, and the problems will carry over to Stage 2. This is an example of the classic endogeneity problem, for which the standard solution is to resort to instrumental variables that are correlated with inputs but uncorrelated with inefficiency (e.g., Greene 2003). The instrumental variables approach might be applicable to the CNLS estimator. Second, if $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) \neq \sigma^2$ (i.e., inefficiencies exhibit heteroskedasticity or serial correlation), then the CNLS estimator used in Stage 1 remains unbiased and consistent, but is likely inefficient. Moreover, inefficiency in Stage 1 estimation will likely cause downward bias in Stage 2. If the covariance matrix $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})$ can be consistently estimated, an efficient generalized least-squares estimator (Greene 2003) could be adapted to CNLS in a relatively straightforward fashion. Third, if inefficiency terms $\boldsymbol{\varepsilon}$ are uncorrelated but the i.i.d. assumption fails, then the CNLS estimator used in Stage 1 will capture the shape of the frontier correctly, but the inefficiency estimates obtained in Stage 2 will likely overestimate the true inefficiency. We test robustness of C^{2}NLS estimators to these types of violations by means of Monte Carlo simulations in §5.3. More in-depth treatment of these types of violations falls beyond the scope of the present paper, and is left as an interesting topic for future research.

The DEA estimator is generally downward biased (e.g., Simar and Wilson 2000). In a small sample, the C^{2}NLS estimator may be biased, but the direction and the magnitude of the bias are difficult to predict. The CNLS estimator used in Stage 1 is unbiased, even in a small sample. However, any estimation error in Stage 1 is likely to cause upward bias in Stage 2. On the other hand, the maximum CNLS residual ($\max_h \varepsilon_h^{CNLS}$) used in Stage 2 is a downward-biased estimator of the expected inefficiency because the most efficient firm in the sample may not be perfectly efficient relative to the true frontier f. In a small sample, the direction and magnitude of bias depends on the number of input variables, curvature of the true frontier, distribution of the sample firms, and other random factors. Of course, one could apply bootstrapping methods for quantifying the possible bias and for correcting for the estimates, directly analogous to DEA (see, e.g., Simar and Wilson 1998, 2000). Bootstrapping can be adapted to C^{2}NLS in a straightforward fashion to draw statistical inferences, including confidence intervals and hypothesis testing. Besides bootstrapping methods, nonparametric statistical tests (such as Kolmogorov-Smirnov) could be applied for hypothesis testing in the C^{2}NLS context (cf. Banker 1993).

As the sample size increases, we can expect the bias of C^{2}NLS to diminish. In fact, Theorem 4.1 implies the following asymptotic result.

COROLLARY TO THEOREM 4.1. *Under the assumptions stated in Theorem* 4.1, *the C^{2}NLS efficiency estimator is asymptotically unbiased, that is,*

$$\lim_{n\to\infty} Bias[\hat{\varepsilon}_i^{C2NLS}] = \lim_{n\to\infty} (E[\hat{\varepsilon}_i^{C2NLS}] - \varepsilon_i) = 0 \quad \forall i = 1, \ldots, n. \quad (14)$$

In practice, the downward bias of DEA estimators can hamper their discriminatory power. For example, in DEA a firm that consumes the smallest amount of any input will appear efficient by construction. However, in C^{2}NLS, the smallest input value does not immediately guarantee efficiency. Typically, there is only a single efficient firm in C^{2}NLS (i.e., firm $h = \arg\max_i\{\varepsilon_i^{CNLS}\}$). Thus, we can prove this formal result.

THEOREM 4.2. *For any real-valued data set, the discriminatory power of C^{2}NLS is always greater than or equal to that of DEA in the sense that*

$$\hat{\varepsilon}_i^{C2NLS} \leqslant \varepsilon_i^{DEA} \leqslant 0 \quad \forall i = 1, \ldots, n. \quad (15)$$

It should be emphasized that the greater discriminatory power of C^{2}NLS does not say anything about the precision of the estimates; Theorem 4.2 only establishes that the C^{2}NLS method will always yield lower efficiency estimates (i.e., higher degree of inefficiency) than the DEA method. This result applies only to the absolute efficiency estimates; the efficiency ranking of an arbitrary firm i may be higher according to the DEA method than according to C^{2}NLS.

We must also emphasize that the C^{2}NLS method assumes the data-generating process to be deterministic; similar to standard DEA or the parametric PP and COLS methods, it is not designed to noisy environments. In contrast to DEA, however, all observations influence the shape of the C^{2}NLS frontier. Thus, a single outlier located above the true frontier does not distort the shape of the C^{2}NLS frontier as severely as in DEA. Further, C^{2}NLS utilizes the information that inefficient observations contain about the frontier. If outliers are a concern, conditional quantile methods (e.g., Daraio and Simar 2007) could also be applied to C^{2}NLS.

In conclusion, C^{2}NLS is a new nonparametric approach to efficiency analysis in the deterministic setting. Besides providing efficiency estimates, the new C^{2}NLS approach can be used for estimating shadow prices, setting performance targets, and identifying benchmarks in a similar fashion as the standard DEA. Many of the established techniques from the DEA toolbox, such as returns to scale modeling, weight restrictions, conditional quantiles, or statistical inferences through bootstrapping, can be directly incorporated in the C^{2}NLS framework as well. The next section discusses the Monte Carlo evidence regarding the comparison of several deterministic frontier estimation methods.

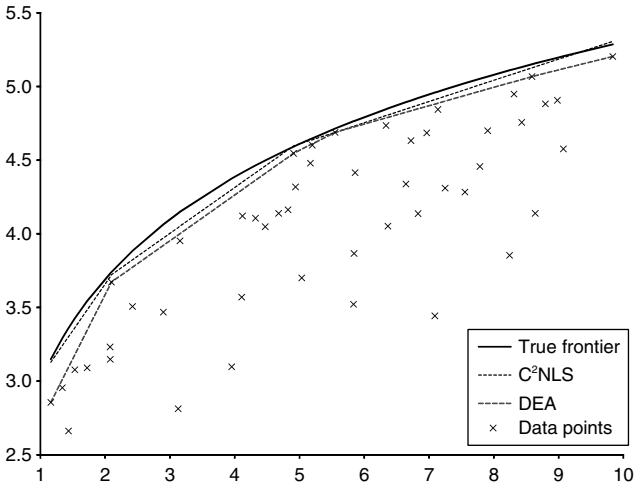
5. Monte Carlo Simulations

This section describes simulation results that serve four purposes: visualization, comparison of methods, investigation of the interaction of the error term, and investigation of the effects of noise on the deterministic estimators. First, we present a graph of frontiers constructed by DEA and C^{2}NLS methods to visualize C^{2}NLS and its relative merits. Second, four frontier methods are assessed based on bias and mean squared error for six different scenarios. Third, we examine the robustness of the frontier-shifting methods, COLS and C^{2}NLS, to misspecification of interaction of the error term with the production function, and consider the additive and multiplicative error terms that are consistent with the model assumptions of the shifting methods. Finally, we investigate the performance of the same four frontier methods when inefficiency is measured in the presence of noise.

5.1. Illustration

This subsection illustrates the frontiers estimated by DEA and C^{2}NLS. For this example, we assume a production function $y = 3 + \ln(x) - u$ and generate 50 observations where x is randomly sampled from a uniform distribution $Uni[1, 10]$ and u is a draw from a half-normal distribution with standard deviation of 0.7. Figure 1 graphically illustrates the data (points x), the true frontier (thick grey curve), the C^{2}NLS frontier (thin black broken piecewise-linear curve), and the DEA frontier (light grey broken piecewise-linear curve). DEA and C^{2}NLS estimate piecewise-linear frontiers with five and three segments, respectively. We find that both estimation methods produce good approximations of the frontier for a production process using only one input to generate one output. The results described in Theorem 4.2 can be seen because the C^{2}NLS frontier has greater than or equal output levels for all input levels. In short, both methods perform well, and C^{2}NLS slightly underestimates the

Figure 1. Graphical illustration of DEA and C^{2}NLS for an example data set.



frontier when averaged across all input levels, whereas DEA underestimates the frontier for all input levels, with a larger average deviation from the frontier.

5.2. Comparison of Frontier Estimation Methods

Next, we present a more systematic comparison of frontier estimation methods in alternative simulated environments. We restricted our analysis to PP, COLS, DEA, and C^{2}NLS considered above. A function linear in inputs is used for PP and COLS. Table 2 describes the six simulated scenarios considered. Scenarios A and B represent a single-input case with different production functions, Scenario C involves two inputs, and Scenario D three inputs. Scenarios E and F adjust the functional forms used in scenarios C and D to increase the curvature in the frontier. For all scenarios, we tested three data set sizes of 50, 100, and 150 observations. The input data were randomly sampled from $Uni[1, 10]$, independently for each input and firm. Then, the efficient output levels were calculated and a random inefficiency term $u \underset{\text{i.i.d.}}{\sim} |N(0, 0.4)|$ was subtracted to obtain the data used for the analysis. We ran 100 trials for each combination of scenario and data set size to investigate the relative performance of the four estimation methods.

Performance of each method is evaluated by two standard criteria: the mean-squared error (MSE) and the bias. The MSE statistic is defined as

$$\text{MSE} = \sum_{t=1}^{M} \sum_{i=1}^{n} (\hat{f}(x_i) - f(x_i))^2 / nM, \tag{16}$$

where $\hat{f}$ indicates the estimated frontier production function identified by the particular method (e.g., DEA), and M is the number of simulation runs (here $M = 100$). Similarly, the bias statistic is calculated as

$$\text{BIAS} = \sum_{t=1}^{M} \sum_{i=1}^{n} (\hat{f}(x_i) - f(x_i)) / nM. \tag{17}$$

Although the bias statistic indicates whether the estimated frontier $\hat{f}$ systematically underestimates (BIAS < 0) or overestimates (BIAS > 0) the true frontier f, we note that positive and negative deviations cancel out when averaged over observations and simulation runs. The MSE statistic measures precision of estimates in quadratic terms, assigning equal weight to both positive and negative deviations.

Table 2. Description of the six scenarios—the additive inefficiency case.

Scenario	Inputs	Functional form
(A)	x	$y = \ln(x) + 3 - u$
(B)	x	$y = 3 + x^{1/2} + \ln(x) - u$
(C)	x_1, x_2	$y = 0.1x_1 + 0.1x_2 + 0.3(x_1x_2)^{1/2} - u$
(D)	x_1, x_2, x_3	$y = 0.1x_1 + 0.1x_2 + 0.1x_3 + 0.3(x_1x_2x_3)^{1/3} - u$
(E)	x_1, x_2	$y = 0.1x_1 + 0.1x_2 + 0.3(x_1x_2)^{1/3} - u$
(F)	x_1, x_2, x_3	$y = 0.1x_1 + 0.1x_2 + 0.1x_3 + 0.3(x_1x_2x_3)^{1/4} - u$

Table 3 reports the MSE and bias statistics for the alternative methods in scenarios A–F. The first two columns indicate the scenario and the sample size. The next four columns indicate the MSE of the four methods considered. The center column reports the fraction of trials in which the MSE of the C^{2}NLS method was smaller than the MSE of the DEA method. The next four columns report the bias statistics for each method. The final column presents the fraction of trials in which the C^{2}NLS method had a smaller bias.

Scenarios A and B are single-input, single-output analysis. Scenario B uses a slightly more complicated generating function, resulting in a frontier that is less steep. However, even with the flatter frontier, the performances of the shifting methods, COLS and C^{2}NLS, are comparable to PP and DEA based on a mean squared error (MSE) or bias measure. The parametric methods are limited by the functional form assumption for the single-input cases. As the dimensionality of the model increases beyond trivial cases, C^{2}NLS consistently outperforms DEA. Also, all methods are affected by the increase in dimensionality from a single input to multiple inputs. However, the parametric assumptions of PP and COLS make these methods more robust to the curse of dimensionality than their nonparametric counterparts. When the number of inputs is increased from two to three for scenarios C and D, the parametric methods maintain similar performance, but the nonparametric models' performance is negatively impacted. As the dimensionality increases and the number of observations decreases, C^{2}NLS outperforms DEA and is competitive with the parametric methods. Both nonparametric methods' performance is maintained as the curvature of the frontier is increased.

5.3. Test for Robustness Against Misspecification of the Error Term Interactions

This section analyzes the same set of production functions when the inefficiency term interacts in a multiplicative manner. The parametric methods (PP, COLS) and the proposed C^{2}NLS method assume that the error terms (the inefficiency term in these deterministic cases) are independently and identically distributed (i.i.d.). The multiplicative error term violates this assumption, causing heteroskedasticity; observations with larger input levels tend to be located further from the frontier when measured in output units. This should favor the DEA estimators that do not require the i.i.d. assumption. We employed the same scenarios, data set sizes, input, and efficiency generation process to investigate the multiplicative interactions. Table 4 describes the six simulation scenarios considered.

The same four estimation methods were evaluated under the multiplicative inefficiency term specification, and the criteria of MSE and bias were again used to quantify the performance. In Table 5, we can observe that DEA outperforms C^{2}NLS on the MSE performance measure for experiments where the number of observations is large and the number of dimensions of the analysis is small. However,

Table 3. Relative performance of estimation methods for the additive error models based on MSE and bias performance criteria.

Scenario	Number of obs.	Mean squared error				Fraction of trials DEA > C^2NLS	Bias				Fraction of trials abs(DEA) > abs(C^2NLS)
		PP	COLS	DEA	C^2NLS		PP	COLS	DEA	C^2NLS	
(A)	50	0.042	0.044	0.009	0.005	0.76	0.105	0.129	−0.068	0.015	0.87
(A)	100	0.044	0.045	0.004	0.004	0.47	0.119	0.139	−0.040	0.034	0.62
(A)	150	0.045	0.047	0.003	0.004	0.49	0.125	0.145	−0.035	0.032	0.62
(B)	50	0.095	0.100	0.011	0.008	0.66	0.177	0.205	−0.074	0.028	0.81
(B)	100	0.101	0.104	0.005	0.006	0.46	0.193	0.218	−0.048	0.043	0.62
(B)	150	0.102	0.105	0.003	0.005	0.36	0.200	0.223	−0.038	0.038	0.52
(C)	50	0.022	0.029	0.032	0.013	0.95	0.053	0.095	−0.130	0.003	0.96
(C)	100	0.023	0.029	0.019	0.011	0.87	0.071	0.103	−0.092	0.039	0.83
(C)	150	0.024	0.030	0.013	0.009	0.72	0.079	0.113	−0.074	0.049	0.68
(D)	50	0.022	0.029	0.070	0.025	0.99	0.046	0.104	−0.207	−0.024	0.99
(D)	100	0.023	0.031	0.046	0.017	0.98	0.076	0.119	−0.162	0.023	0.97
(D)	150	0.025	0.033	0.035	0.015	0.97	0.092	0.129	−0.138	0.049	0.97
(E)	50	0.005	0.007	0.030	0.015	0.91	0.001	0.035	−0.121	0.018	0.94
(E)	100	0.004	0.006	0.016	0.011	0.85	0.019	0.041	−0.083	0.044	0.80
(E)	150	0.004	0.005	0.011	0.008	0.72	0.026	0.046	−0.068	0.046	0.71
(F)	50	0.008	0.010	0.061	0.022	0.99	−0.003	0.049	−0.191	−0.015	0.99
(F)	100	0.006	0.009	0.041	0.017	0.95	0.023	0.060	−0.148	0.038	0.94
(F)	150	0.006	0.009	0.030	0.013	0.94	0.034	0.064	−0.124	0.043	0.92

even with the misspecification of the inefficiency interaction, C^2NLS dominates DEA for higher dimensionality models. Again in higher dimensions, the relative robustness of PP and COLS to the curse of dimensionality can be observed.

On the measure of bias, the specification error increases the positive bias of C^2NLS in most scenarios. Nevertheless, the absolute bias of C^2NLS is on average smaller than that of DEA. Unlike the analysis of the additive production functions, PP outperforms the other models on the basis of bias for most scenarios.

Overall, the simulation analysis reveals that C^2NLS is likely to perform better than DEA as the ratio of observations to dimensionality of the model decreases. And, as the complexity of production processes increases, involving multiple inputs, the superiority of C^2NLS becomes more pronounced. The curse of dimensionality is observed in this analysis because the nonparametric methods performance deteriorates more quickly than the parametric methods as the dimensionality of the models grows. Also, we observe that whereas misspecification of the interaction of the inefficiency component worsens the performance of the frontier-shifting methods (COLS, C^2NLS), the same result, C^2NLS outperforming DEA as the ratio of observations to dimensionality of the model decreases, is observed. As the curvature of the frontier is increased, the performance of C^2NLS decreases slightly, but still continues to outperform DEA. This makes a strong case for C^2NLS as an alternative method to consider when deciding among nonparametric methods, particularly when realistic model sizes and data availability are considered.

Table 4. Description of the six scenarios—the multiplicative inefficiency case.

Scenario	Inputs	Functional form
(A)	x	$y = (\ln(x) + 3)/(1 + u)$
(B)	x	$y = (3 + x^{1/2} + \ln(x))/(1 + u)$
(C)	x_1, x_2	$y = (0.1x_1 + 0.1x_2 + 0.3(x_1x_2)^{1/2})/(1 + u)$
(D)	x_1, x_2, x_3	$y = (0.1x_1 + 0.1x_2 + 0.1x_3 + 0.3(x_1x_2x_3)^{1/3})/(1 + u)$
(E)	x_1, x_2	$y = (0.1x_1 + 0.1x_2 + 0.3(x_1x_2)^{1/3})/(1 + u)$
(F)	x_1, x_2, x_3	$y = (0.1x_1 + 0.1x_2 + 0.1x_3 + 0.3(x_1x_2x_3)^{1/4})/(1 + u)$

5.4. Estimating Frontiers in the Presence of Noise

In this subsection, we examine the robustness of PP, COLS, DEA, and C^2NLS methods to stochastic noise in data. We emphasize that none of these methods are designed for noisy environments, and none of these methods make any attempt to quantify the magnitude of noise or adjust the efficiency estimates for it. If the practitioner believes that noise is a significant issue, we would not recommend using any of these four methods, but rather choosing parametric SFA or nonparametric StoNED (Kuosmanen and Kortelainen 2007). However, if the data are free from noise, the four deterministic methods considered are likely to be more efficient, in a statistical sense, than SFA or StoNED. In practice, it is difficult to test whether or not data are perturbed by noise. Therefore, robustness to noise is a useful property even for deterministic methods that are not explicitly designed for dealing with noise.

To generate noisy data, we modify scenario D of §5.2 by adding a random disturbance term v to obtain

$$y = f(\mathbf{x}) - u + v, \tag{18}$$

Table 5. Relative performance of estimation methods for the multiplicative models based on MSE and bias performance criteria.

Scenario	Number of obs.	Mean squared error				Fraction of trials DEA > C²NLS	Bias				Fraction of trials abs(DEA) > abs(C²NLS)
		PP	COLS	DEA	C²NLS		PP	COLS	DEA	C²NLS	
(A)	50	0.049	0.054	0.009	0.007	0.67	0.021	0.064	−0.069	0.024	0.82
(A)	100	0.045	0.054	0.004	0.005	0.44	0.057	0.093	−0.044	0.035	0.59
(A)	150	0.046	0.057	0.003	0.004	0.46	0.083	0.110	−0.035	0.034	0.55
(B)	50	0.115	0.156	0.011	0.008	0.75	0.035	0.116	−0.075	0.026	0.83
(B)	100	0.104	0.160	0.005	0.006	0.41	0.088	0.166	−0.048	0.041	0.59
(B)	150	0.108	0.167	0.004	0.005	0.41	0.127	0.191	−0.039	0.040	0.53
(C)	50	0.030	0.133	0.031	0.016	0.93	0.005	0.230	−0.126	0.029	0.94
(C)	100	0.025	0.149	0.019	0.010	0.87	0.040	0.280	−0.093	0.036	0.83
(C)	150	0.024	0.171	0.013	0.010	0.71	0.057	0.319	−0.074	0.058	0.62
(D)	50	0.044	0.118	0.065	0.025	0.98	−0.040	0.192	−0.201	−0.019	0.98
(D)	100	0.029	0.159	0.046	0.019	0.98	0.022	0.293	−0.161	0.038	0.98
(D)	150	0.026	0.169	0.034	0.015	0.96	0.045	0.320	−0.138	0.045	0.95
(E)	50	0.009	0.041	0.204	0.076	0.95	−0.024	0.113	−0.346	0.009	0.95
(E)	100	0.005	0.042	0.123	0.062	0.83	0.003	0.137	−0.253	0.101	0.80
(E)	150	0.005	0.049	0.090	0.076	0.69	0.013	0.160	−0.204	0.159	0.64
(F)	50	0.023	0.055	0.529	0.150	1.00	−0.060	0.112	−0.590	−0.097	1.00
(F)	100	0.010	0.069	0.338	0.116	0.97	−0.012	0.180	−0.452	0.087	0.96
(F)	150	0.008	0.070	0.249	0.124	0.88	0.005	0.196	−0.374	0.171	0.85

where

$$f(\mathbf{x}) = 0.1x_1 + 0.1x_2 + 0.1x_3 + 0.3(x_1x_2x_3)^{1/3}. \quad (19)$$

The data-generating process of u and v terms is controlled through parameters $\tilde{\lambda} = \sigma_u/\sigma_v$ (representing the signal-to-noise ratio, not to be confused with the intensity weights of DEA) and $\sigma^2 = \sigma_u^2 + \sigma_v^2$, where σ_u^2, σ_v^2 are the variances of u and v, respectively. Setting $\sigma^2 = 0.2 \cdot \mathrm{Var}[f(\mathbf{x})]$, we consider five different levels of $\tilde{\lambda}$. The first values 0.83 and 1.66 have been adopted from Aigner et al. (1977) to represent very noisy environment (note: When $\tilde{\lambda} < 1$, the noise term v has a larger variance than the inefficiency term u). We then increase the value of $\tilde{\lambda}$ in equal steps until value 4.15, which represents a low-noise setting.

Table 6 reports the results from analyzing all four methods for varying levels of $\tilde{\lambda}$. These results indicate that all methods perform worse in the presence of noise, both in terms of MSE and bias. Note that all four methods overestimate the frontier in the case of heavy noise: Bias is positive when $\tilde{\lambda} = 0.83$. Interestingly, the nonparametric approaches consistently outperform the parametric counterparts. Good performance of DEA is largely explained by the fact that the outliers generated by the noise term v (which cause positive bias to all four methods) serve to offset the small sample bias of DEA (i.e., the systematic negative bias observed in Tables 3 and 5). This offsetting effect also explains the low MSE of DEA. As $\tilde{\lambda}$ grows larger (low-noise scenarios), the small sample-bias dominates, and C²NLS outperforms DEA in both MSE and bias.

The MSE and bias statistics refer to the estimated versus true frontier. In the stochastic setting, good performance in estimating the frontier does not yet guarantee that the efficiency estimates and efficiency rankings are robust to noise. To assess robustness of efficiency rankings, the rank correlations between the estimated and true inefficiency terms u were computed for DEA and C²NLS. These results are presented in Table 7. We observe that C²NLS outperforms DEA by rank correlation in virtually all cases. On average, the efficiency rankings obtained with C²NLS are

Table 6. Performance of four estimation methods as measured by MSE and bias performance criteria in a scenario with noise [(18), (19)], sample size $n = 100$.

Lambda	Scenario									
	Mean squared error (MSE)				Fraction of trials DEA > C²NLS	Bias				Fraction of trials abs(DEA) > abs(C²NLS)
	PP	COLS	DEA	C²NLS		PP	COLS	DEA	C²NLS	
0.83	0.33	0.64	0.08	0.27	0.00	0.53	0.75	0.12	0.48	0.00
1.66	0.15	0.32	0.04	0.09	0.1	0.34	0.51	−0.02	0.26	0.01
2.49	0.09	0.25	0.04	0.05	0.42	0.24	0.45	−0.08	0.17	0.2
3.32	0.06	0.2	0.04	0.03	0.78	0.2	0.39	−0.12	0.1	0.57
4.15	0.06	0.18	0.04	0.03	0.85	0.18	0.37	−0.13	0.09	0.72

Table 7. Performance of DEA and C^{2}NLS as measured by rank correlation, the noisy scenario (18), sample size $n = 100$.

	Scenario		
	Rank correlation		
Lambda	DEA	C^{2}NLS	Fraction of trials DEA > C^{2}NLS
0.83	0.26	0.37	0.04
1.66	0.43	0.58	0.00
2.49	0.50	0.70	0.00
3.32	0.55	0.76	0.00
4.15	0.59	0.79	0.00

more closely correlated with the true efficiency rankings than the rankings obtained with DEA. Although the C^{2}NLS efficiency estimates are likely to underestimate the true efficiency levels due to outliers (causing the positive bias in Table 6), the efficiency rankings from C^{2}NLS are relatively robust to small and medium-sized noise.

In conclusion, the presented simulation results indicate that the proposed C^{2}NLS method is a competitive alternative for the existing parametric and nonparametric techniques when there is little or no noise in the data. As evident from the simulations, the main advantage of C^{2}NLS compared to DEA is its robustness to small-sample error, which is particularly beneficial when the sample size is small compared to the number of inputs (the curse of dimensionality). We have also examined robustness of C^{2}NLS to heteroskedasticity and stochastic noise. If the variance of noise is relatively small compared to that of inefficiency, C^{2}NLS can yield more robust efficiency rankings than DEA. If the data is expected to be very noisy, methods that explicitly account for noise (such as SFA or StoNED) would be preferred.

6. Conclusions

We have presented a new least-squares interpretation of the DEA model. It was shown that DEA can be recast as nonparametric least-squares regression subject to shape constraints on the frontier and sign constraints on residuals. This connection between DEA and least-squares regression further contributes to developing the statistical foundation of DEA, opening up new ways for adapting and integrating econometric techniques to DEA. Next, the parallel development of parametric and nonparametric models was outlined. We showed that Aigner and Chu's (1968) parametric programming model is a constrained special case of DEA in which a functional form for the production function is assumed. The parallel development of parametric and nonparametric models was further extended by the introduction of C^{2}NLS and the linkage between COLS and C^{2}NLS was established.

Performance of the new C^{2}NLS method was investigated via Monte Carlo simulation. The results indicated that C^{2}NLS performs at least as well as DEA in cases when the data-to-dimension ratio is relatively high, and significantly outperforms DEA as this ratio falls. Both additive and multiplicative error specifications were considered, with C^{2}NLS proving robust to either specification. Consistency and asymptotic unbiasedness of the C^{2}NLS efficiency estimator are established, but the properties of the C^{2}NLS estimator would clearly warrant further research. In contrast to DEA, the C^{2}NLS frontier uses the information contained in both efficient and inefficient observations, and thus is expected to be less sensitive to outliers and extreme observations.

We believe that our findings, combining insights from the econometric and operations research domains, promote the development of a unified framework for productive efficiency analysis. The results we have described create the foundation for the development of a truly nonparametric stochastic efficiency model. By linking the nonparametric frontier estimation methods to regression, an error term with a stochastic component can be specified, similar to the stochastic frontier model of Aigner et al. (1977). Applying these results to the general multioutput setting provides a challenging opportunity for future research.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnotes

1. Interestingly, Hildreth illustrated his method by estimating production function for cotton production using data from field experiments. While Hildreth focused on estimating the neoclassical production function rather than the best-practice frontier, his nonparametric approach based on shape constraints (monotonicity and concavity) is directly analogous to Farrell (1957) and Afriat (1972). Although there is no evidence that Hildreth's (1954) article inspired subsequent work by Farrell, Afriat, and others, the results established in this paper imply that Clifford Hildreth deserves to be recognized as one of the predecessors to the modern DEA.
2. Afriat's Theorem has been analogously applied for nonparametric estimation in Banker and Maindiratta (1992) and Matzkin (1994).
3. Simulation results investigating the number of hyperplanes generated can be found in Kuosmanen (2008).
4. We liberally use both sum operators and scalar products in the same equations because strict adherence to either convention would make the notation of this paper unnecessarily cryptic.
5. See Kalvelagen (2004) for a more complete discussion of computational issues related to similar formulations.

References

Afriat, S. N. 1967. The construction of a utility function from expenditure data. *Internat. Econom. Rev.* **8** 67–77.
Afriat, S. N. 1972. Efficiency estimation of production functions. *Internat. Econom. Rev.* **13**(3) 568–598.

Aigner, D., S. Chu. 1968. On estimating the industry production function. *Amer. Econom. Rev.* **58** 826–839.

Aigner, D., C. A. K. Lovell, P. Schmidt. 1977. Formulation and estimation of stochastic frontier production function models. *J. Econom.* **6** 21–37.

Banker, R. D. 1993. Maximum-likelihood, consistency and data envelopment analysis—A statistical foundation. *Management Sci.* **39**(10) 1265–1273.

Banker, R. D., A. Maindiratta. 1992. Maximum likelihood estimation of monotone and concave production frontiers. *J. Productivity Anal.* **3** 401–415.

Banker, R. D., A. Charnes, W. W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Sci.* **30**(9) 1078–1092.

Banker, R. D., S. M. Datar, C. F. Kemerer. 1991. A model to evaluate variables impacting the productivity of software maintenance projects. *Management Sci.* **37**(1) 1–18.

Charnes, A., W. W. Cooper, E. Rhodes. 1978. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **2** 429–444.

Cobb, C. W., P. H. Douglas. 1928. A theory of production. *Amer. Econom. Rev.* **18** 139–165.

Cooper, W. W., Z. M. Huang, S. X. Li. 1996. Satisficing DEA models under chance constraints. *Ann. Oper. Res.* **66** 279–295.

Cooper, W. W., L. M. Seiford, J. Zhu. 2004. Data envelopement analysis: Models and interpretations. W. W. Cooper, L. M. Seiford, J. Zhu, eds. *Handbook on Data Envelopment Analysis*, Chapter 1. Kluwer Academic Publishers, Boston, 1–39.

Daraio, C., L. Simar. 2007. Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. *J. Productivity Anal.* **28**(1–2) 13–32.

Farrell, M. J. 1957. The measurement of productive efficiency. *J. Royal Statist. Soc. Ser. A. Statist. Soc.* **120**(3) 253–281.

Fraser, D. A. S., H. Massam. 1989. A mixed primal-dual based algorithm for regression under inequality constraints: Application to convex regression. *Scandinavian J. Statist.* **16** 65–74.

Gattoufi, S., M. Oral, A. Kumar, A. Reisman. 2004. Content analysis of data envelopment analysis literature and its comparison with that of other OR/MS fields. *J. Oper. Res. Soc.* **55**(9) 911–935.

Greene, W. 1980. Maximum likelihood estimation of econometric frontier functions. *J. Econometrics* **13** 26–57.

Greene, W. H. 2003. *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ.

Groeneboom, P., G. Jongbloed, J. A. Wellner. 2001. A canonical process for estimation of convex functions: The "invelope" of integrated Brownian motion plus $t(4)$. *Ann. Statist.* **29**(6) 1620–1652.

Gstach, D. 1998. Another approach to data envelopment analysis in noisy environments: DEA+. *J. Productivity Anal.* **9**(2) 161–176.

Hanson, D. L., G. Pledger. 1976. Consistency in concave regression. *Ann. Statist.* **4**(6) 1038–1050.

Hildreth, C. 1954. Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49**(267) 598–619.

Kalvelagen, E. 2004. Efficiently solving DEA models with GAMS. GAMS, Washington, DC, 22.

Korostelev, A., L. Simar, B. Tsybakov. 1995. On estimation of monotone and convex boundaries. *Publ. Inst. Statist. University of Paris* **XXXIX**(1) 3–18.

Kuosmanen, T. 2006. Stochastic nonparametric envelopment of data: Combining virtues of SFA and DEA in a unified framework. MTT Discussion papers, Helsinki, Finland.

Kuosmanen, T. 2008. Representation theorem for convex nonparametric least squares. *Econometrics J.* **11** 308–325.

Kuosmanen, T., M. Fosgerau. 2009. Neoclassical versus frontier production models? Testing for the presence of inefficiencies in the regression residuals. *Scandinavian J. Econom.* **111**(2) 317–333.

Kuosmanen, T., M. Kortelainen. 2007. Stochastic nonparametric envelopment of data: Cross-sectional frontier estimation subject to shape constraints. Economics discussion paper 46, University of Joensuu, Joensuu, Finland.

Kuosmanen, T., L. Cherchye, T. Sipilainen. 2006. The law of one price in data envelopment analysis: Restricting weight flexibility across firms. *Eur. J. Oper. Res.* **170**(3) 735–757.

Mammen, E. 1991. Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

Mammen, E., C. Thomas-Agnan. 1999. Smoothing splines and shape restrictions. *Scandinavian J. Statist.* **26** 239–252.

Matzkin, R. L. 1994. Restrictions of economic theory in nonparametric methods. R. F. Engle, D. L. McFadden, eds. *Handbook of Econometrics*, Vol. IV. Elsevier, Amsterdam, 2523–2558.

Meeusen, W., J. Van den Broeck. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *Internat. Econom. Rev.* **18**(2) 435–445.

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

Nemirovskii, A. S., B. T. Polyak, A. B. Tsybakov. 1985. Rates of convergence of nonparametric estimates of maximum likelihood type. *Problems Inform. Transmission* **21** 258–271.

Schmidt, P. 1976. On the statistical estimation of parametric frontier production functions. *Rev. Econom. Statist.* **58** 238–239.

Schmidt, P. 1985. Frontier production functions. *Econometric Rev.* **4**(2) 289–328.

Seiford, L. M. 1996. Data envelopment analysis: The evolution of the state of the art (1978–1995). *J. Productivity Anal.* **7**(2–3) 99–137.

Simar, L., P. W. Wilson. 2000. A general methodology for bootstrapping in non-parametric frontier models. *J. Appl. Statist.* **27**(6) 779–802.

Simar, L., P. W. Wilson. 2008. Statistical inference in nonparametric frontier models: Recent developments and perspectives. H. O. Fried, C. A. K. Lovell, S. S. Schmidt, eds. *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, New York, 421–451.

Stone, C. J. 1980. Optimal rates of conference for nonparametric estimators. *Ann. Statist.* **8**(6) 1348–1360.

Timmer, C. P. 1971. Using a probabilistic frontier production function to measure technical efficiency. *J. Political Econom.* **79** 767–794.

Winsten, C. B. 1957. Discussion on Mr. Farrell's paper. *J. Royal Statist. Soc. Ser. A. Statist. Soc.* **120**(3) 282–284.

Yatchew, A. 1998. Nonparametric regression techniques in economics. *J. Econom. Literature* **36**(2) 669–721.

Yatchew, A. 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, New York.