Stochastic nonparametric approach to efficiency analysis:

A Unified Framework

Timo Kuosmanen[1], Andrew Johnson[2,1], Antti Saastamoinen[1]

1) School of Business, Aalto University, 00100 Helsinki, Finland.

2) Department of Industrial and Systems Engineering, Texas A&M University, TX 77840, USA.

Abstract

Bridging the gap between axiomatic *Data Envelopment Analysis* (DEA) and econometric *Stochastic Frontier Analysis* (SFA) has been one of the most vexing problems in the field of efficiency analysis. Recent developments in multivariate convex regression, particularly *Convex Nonparametric Least Squares* (CNLS) method, have led to the full integration of DEA and SFA into a unified framework of productivity analysis, referred to as *Stochastic Nonparametric Envelopment of Data* (StoNED). The unified framework of StoNED offers a general and flexible platform for efficiency analysis and related themes such as frontier estimation and production analysis, allowing one to combine existing tools of efficiency analysis in novel ways across the DEA-SFA spectrum, facilitating new opportunities for further methodological development. This chapter provides an updated and elaborated presentation of the CNLS and StoNED methods. This chapter also extends the scope of the StoNED method in several directions. Most notably, this chapter examines quantile estimation using StoNED and an extension of the StoNED method to the general case of multiple inputs and multiple outputs. This chapter also provides a detailed discussion of how to model heteroscedasticity in the inefficiency and noise terms.

Key words:

Efficiency analysis, Frontier estimation, Multivariate convex regression, Nonparametric least squares, Productivity, Stochastic noise.

December 2013, revised June 2014

Forthcoming as a chapter in

J. Zhu (Ed.) *Handbook on Data Envelopment Analysis Vol II,* Springer.

1. Introduction

Efficiency analysis is an essential and extensive research area that provides answers to such important questions as: Who are the best performing firms and can we learn something from their behavior? [1] What are the sources of efficiency differences across firms? Can efficiency be improved by government policy or better managerial practices? Are there benefits to increasing the scale of operations? These are examples of important questions we hope to resolve with efficiency analyses.

Efficiency analysis is an interdisciplinary field that spans such disciplines as economics, econometrics,[2] operations research and management science,[3] and engineering, among others. The methods of efficiency analysis are utilized in several fields of application including agriculture, banking, education, environment, health care, energy, manufacturing, transportation, and utilities, among many others. Efficiency analysis is performed at various different scales. Micro level applications range from individual persons, teams, production plants and facilities to company level and industry level efficiency assessments. Macro level applications range from comparative efficiency assessments of production systems or industries across countries to efficiency assessment of national economies. Indeed, efficiency improvement is one of the key components of productivity growth (e.g., Färe et al., 1994), which in turn is the primary driver of economic welfare. The benefits to understanding the relationship between efficiency and productivity and quantifying efficiency cannot be overstated. In words of Paul Krugman (1992, p. 9), "*Productivity isn't everything, but in the long run it is almost everything. A country's ability to improve its standard of living over time depends almost entirely on its ability to raise its output per worker*." Note that macro-level performance of a country is an aggregate of the individual firms operating within that country. Therefore, sound micro-foundations of efficiency analysis are critical for the integrity of productivity and efficiency analysis at macro level.

Unfortunately, there currently is no commonly accepted methodology of efficiency analysis, but the field is divided between two competing approaches: Data envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA). [4]

Data envelopment analysis (DEA, Farrell, 1957; Charnes et al., 1978) is an axiomatic, mathematical programming approach to efficiency analysis. DEA's main advantage compared to econometric, regression-based tools is its nonparametric treatment of the frontier, building upon axioms of production theory such as free disposability (monotonicity), convexity (concavity), and constant returns to scale (homogeneity). DEA does not assume any particular functional form for the frontier or the distribution of inefficiency. It's direct, data-driven approach is helpful for communicating the results of efficiency analysis to decision-makers. However, the main shortcoming of DEA is that it attributes all deviations from the frontier to inefficiency. This is often a heroic assumption.

[1] We will henceforth use the term "firm" referring to any production unit that transforms inputs to output, including both non-profit and for-profit organizations. The firm can refer to an establishment (facility) or sub-division of a company or to an aggregate entity such as an industry, a region, or a country.

[2] Observe that 13 of the 100 most cited articles published in a leading field journal, the *Journal of Econometrics*, are efficiency analysis papers, including Simar and Wilson (2007) that has 436 citations, making it the #32 most cited paper in the journal in just 6 years from its publication (citations data gathered from Scopus, Nov 25, 2013).

[3] In operations research and management science, Charnes et al. (1978) ranks #1 as most cited article published in the *European Journal of Operational Research* (EJOR) and Banker et al. (1984) is the #1 most cited article in *Management Science*, two of the leading journals of this field (the flagship journals of EURO and INFORMS, respectively). In fact, Charnes et al. article has more than 5 times more citations than the 2nd most cited paper in EJOR (Nov 25, 2013).

[4] Citation statistics of some of the key papers provide undisputable evidence about the significant influence of this field. The four most cited papers are Charnes et al. (1978) with 6,152 citations, Banker et al. (1984) with 3,415 citations, Farrell (1957) with 3,296 citations, and Aigner et al. (1977) with 1,875 citations (Scopus, Nov 25, 2013).

Stochastic frontier analysis (SFA, Aigner, Lovell, Schmidt, 1977; Meeusen and Vanden Broeck, 1977) is often, incorrectly, viewed as a direct competitor of DEA. The key strength of SFA is its probabilistic modeling of deviations from the frontier, which are decomposed into a non-negative inefficiency term and an idiosyncratic error term that accounts for omitted factors such as unobserved heterogeneity of firms and their operating environments, random errors of measurement and data processing, specification errors, and other sources of noise. In contrast to DEA, SFA utilizes parametric regression techniques, which require *ex ante* specifications of the functional forms of the frontier and the inefficiency distribution. Since the economic theory rarely justifies a particular functional form, flexible functional forms such as translog are frequently used. However flexible functional forms often violate axioms of production theory, whereas imposing the axioms will reduce flexibility. In summary, the DEA and SFA methods are not direct competitors but rather complements: in the tradeoff between DEA and SFA something is sacrificed for something to be gained. Namely DEA does not model noise, but is able to impose axiomatic properties and estimate the frontier non-parametrically, while SFA cannot impose axiomatic properties, but has the benefit of modeling inefficiency and noise.

Bridging the gap between axiomatic DEA and stochastic SFA was for a long time one of the most vexing problems in the field of efficiency analysis. The recent works on convex nonparametric least squares (CNLS) by Kuosmanen (2008), Kuosmanen and Johnson (2010), and Kuosmanen and Kortelainen (2012) have led to the full integration of DEA and SFA into a unified framework of productivity analysis, which we refer to as *stochastic nonparametric envelopment of data* (StoNED).[5]

We see the development of StoNED as a paradigm shift for efficiency analysis. It is no longer necessary to decide if modeling noise is more important than imposing axioms of production theory: we can do both using StoNED. The unified framework of StoNED offers deeper insights to the foundations of DEA and SFA, but it also provides a more general and flexible platform for efficiency analysis and related themes such as frontier estimation and production analysis. Further, a number of extensions to the original DEA and SFA methods have been developed over the past decades. The unified StoNED framework allows us to combine the existing tools of efficiency analysis in novel ways across the DEA-SFA spectrum, facilitating new opportunities for further methodological development.

The main objective of this chapter is to provide an updated and elaborated presentation of the CNLS and StoNED methods, the most promising new tools for axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise. Our secondary objective is to extend the scope of the StoNED method in several dimensions. This chapter provides the first extension of the StoNED method to the general case of multiple inputs and multiple outputs. We also consider quantile estimation using StoNED, and present a detailed discussion of how to model heteroscedasticity in the inefficiency and noise terms.

The rest of this chapter is organized as follows. Section 2 introduces the unified StoNED framework and its special cases by reviewing alternative sets of assumptions that motivate different estimation methods applied in productivity analysis. Our focus is explicitly on the axiomatic DEA-style approaches. Section 3 presents the CNLS regression as a quadratic programming problem. Section 4 discusses the intimate connections between CNLS and DEA, and introduces a step-wise C^{2}NLS estimator. Section 5 further develops the step-wise estimation approach for the StoNED

[5] The term StoNED was coined by Kuosmanen (2006). By request of referees, Kuosmanen and Kortelainen (2012) used the term stochastic “non-smooth” envelopment, as their model specification involves parametric distributional assumptions. In this chapter we show that the distributional assumptions can be relaxed: see Sections 5.2.3 and 6.2.

estimator. Section 6 reviews some important extensions to the StoNED, including the multiplicative formulation (Section 6.1), observations from multiple time periods that make up a panel data (Section 6.2), directional distance functions (DDF) for modeling multiple output variables (Section 6.3), and quantile regression formulation (Section 6.4). The model of contextual variables that represent operational conditions or practices is examined in detail in Section 7. Testing of heteroscedasticity and modeling heteroscedasticity of inefficiency and noise using a doubly-heteroscedastic model discussed in Section 8. Finally, Section 9 concludes with discussion of some promising avenues of future research.

2. Unified frontier model

To maintain direct contact with the SFA literature, we introduce the unified model of frontier production function in the multiple input, single output case. Multiple outputs can be modeled using cost functions (see Kortelainen and Kuosmanen, 2012, Section 4.4; and Kuosmanen, 2012) and distance functions. A general multi-input multi-output directional distance function model will be introduced in Section 6.3.

Production technology is represented by a frontier *production function* $f(\mathbf{x})$, where $\mathbf{x}$ is a m-dimensional input vector.[6] Frontier $f(\mathbf{x})$ indicates the maximum output that can be produced with inputs $\mathbf{x}$, and hence the function $f(\mathbf{x})$ characterizes the boundary of the production possibility set. We assume that function f belongs to the class of continuous, monotonic increasing, and globally concave functions that can be non-differentiable (we denote this class as F_2). This is equivalent to stating that the production possibility set satisfies the classic DEA assumptions of free disposability and convexity. In contrast to SFA, no specific functional form for f is assumed.

The observed output y_i of firm i may differ from $f(\mathbf{x}_i)$ due to inefficiency and noise. We follow the SFA literature and introduce a composite error term $\varepsilon_i = v_i - u_i$, which consists of the inefficiency term $u_i > 0$ and the stochastic noise term v_i, formally,

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \varepsilon_i \\ &= f(\mathbf{x}_i) - u_i + v_i, \quad i = 1,...,n \end{aligned} \tag{1}$$

Variables u_i and v_i ($i = 1,...,n$) are random variables that are assumed to be statistically independent of each other as well as of inputs $\mathbf{x}_i$. We assume that the inefficiency term has a positive mean and a constant finite variance, that is, $E(u_i) = \mu > 0$ and $Var(u_i) = \sigma_u^2 < \infty$. We further assume zero mean noise with a constant finite variance, that is, $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2 < \infty$. Assuming σ_u^2 and σ_v^2 are constant across firms is referred to as homoscedasticity; models with heteroskedastic inefficiency and noise will be discussed in Section 8. For the sake of generality and to maintain the fully nonparametric orientation, we do not introduce any distributional assumptions for u_i or v_i at this point. However, some estimation techniques to be introduced below require additional parametric assumptions.

In model (1), the deterministic part (i.e., production function f) is defined analogous to the DEA literature, while the stochastic part (i.e., composite error term ε_i) is defined similar to SFA. As a result, model (1) encompasses the classic models of the SFA and DEA literature as its constrained special cases. Note that in this chapter we use the term “model” in the sense of the

[6] For clarity, we denote vectors by bold lower case letters (e.g., $\mathbf{x}$) and matrices by bold capital letters (e.g., $\mathbf{Z}$). All vectors are column vectors, unless otherwise indicated. Note: $\mathbf{x}'$ denotes the transpose of vector $\mathbf{x}$.

econometric literature to refer to the description of the data generating process (DGP). DEA and SFA are alternative estimators or methods for estimating the production function *f*, the expected inefficiency μ, and the firm-specific realizations of the random inefficiency term u_i. We note that in the DEA literature it is common to use the term "model" for the linear programming problem (e.g., LP model) or other mathematical programming formulations for computing the estimator. To avoid confusion, we will follow the econometric terminology and refer to equation (1) and the related assumptions as the model, whereas DEA, SFA, CNLS, and StoNED are referred to as estimators. In this terminology, "DEA model" or "SFA model" refer to the specific assumptions regarding the variables of model (1).

The literature of efficiency analysis has conventionally focused on fully parametric or nonparametric versions of model (1). Parametric models postulate a priori a specific functional form for *f* (e.g., Cobb-Douglas, translog, etc.) and subsequently estimate its unknown parameters. In contrast, axiomatic nonparametric models assume that *f* satisfies certain regularity axioms (e.g., monotonicity and concavity), but no particular functional form is assumed. At this point, we must emphasize that the term nonparametric does not necessarily imply that there are no restrictive assumptions. It is not true that the assumptions of a nonparametric model are necessarily less restrictive than those of a parametric model. For example, the fully nonparametric DEA estimator of model (1) is based on the assumption of no noise (i.e., $v_i = 0$ for all firms *i*). Assuming away noise does not require any specific parametric specification, but it is nevertheless a restrictive assumption. In fact, it is less restrictive to impose parametric structure and assume v_i are identically and independently distributed according to the normal distribution $N(0,\sigma_v^2)$. Note that this parametric specification contains the fully nonparametric "deterministic" case of no noise as its restricted special case, obtained by imposing the parameter restriction $\sigma_v^2 = 0$.

In addition to the pure parametric and nonparametric alternatives, the intermediate cases of semiparametric and semi-nonparametric models have become increasingly popular in recent years. However, the exact meaning of this terminology is often confused. Chen (2007) provides an intuitive and useful definition that we find worth quoting:

"An econometric model is termed "*parametric*" if all of its parameters are in finite dimensional parameter spaces; a model is "*nonparametric*" if all of its parameters are in infinite-dimensional parameter spaces; a model is "*semiparametric*" if its parameters of interests are in finite-dimensional spaces but its nuisance parameters are in infinite-dimensional spaces; a model is "*semi-nonparametric*" if it contains both finite-dimensional and infinite-dimensional unknown parameters of interests". Chen (2007), p. 5552, footnote 1.

Note that according to the above definition both the semiparametric and semi-nonparametric model contain a nonparametric part and a parametric part. The distinction between the terms semiparametric and semi-nonparametric is subjective, dependent on whether we are interested in the empirical estimates of the nonparametric part or not. The same model can be either semiparametric, if our main interest is in the parameter estimates of the parametric part and the nonparametric part is of no particular interest, or semi-nonparametric, if we are interested in the results of the nonparametric part.

Model (1) can be interpreted as a neoclassical or frontier model depending on the interpretation of the disturbance term (cf., Kuosmanen and Fosgerau, 2009). The neoclassical model assumes that all firms are efficient and disturbances are random, uncorrelated noise terms. Frontier models typically assume that all or some part of the deviations from the frontier are attributed to systematic inefficiency.

Table 1 combines the criteria described above to identify six alternative estimation methods commonly used for estimating the variants of the unified model (1), together with some canonical references. On the parametric side, OLS refers to *ordinary least squares*, PP means *parametric programming*, COLS is *corrected ordinary least squares*, and SFA is *stochastic frontier analysis* (see, e.g., Kumbhakar and Lovell, 2000, for an introduction to the parametric approach to efficiency analysis). The focus of this chapter is on the axiomatic nonparametric and semi-nonparametric variants of model (1): CNLS refers to *convex nonparametric least squares* (Section 3), DEA is *data envelopment analysis* (Section 4.1), C^{2}NLS is *corrected convex non-parametric least squares* (Section 4.2), and StoNED is *stochastic nonparametric envelopment of data* (Section 5).

Table 1. Classification of methods

		Parametric	**Nonparametric**
Central tendency		*OLS* Cobb and Douglas (1928)	*CNLS* (Section 3) Hildreth (1954) Hanson and Pledger (1976)
Deterministic frontier	**Sign constraints**	*PP* Aigner and Chu (1968) Timmer (1971)	*DEA* (Section 4.1) Farrell (1957) Charnes et al. (1978)
	2-step estimation	*COLS* Winsten (1957) Greene (1980)	*C^{2}NLS* (Section 4.2) Kuosmanen and Johnson (2010)
Stochastic frontier		*SFA* Aigner et al. (1977) Meeusen and Vanden Broeck (1977)	*StoNED* (Section 5) Kuosmanen and Kortelainen (2012)

3. Convex nonparametric least squares

In this section we consider the special case of model (1) where the composite error term ε consists exclusively of noise v, and there is no inefficiency (i.e., we assume $u = 0$). This special case is relevant for modeling firms that operate in the competitive market environment, which meets (at least by approximation) the conditions of perfect competition considered in microeconomic theory. We will relax this no inefficiency assumption from Section 4 onwards, but the insights gained in this section will be critical for understanding the developments in the following sections.

In the case of a symmetric zero-mean error term that satisfies $E(\varepsilon_i) = 0$ for all i, the expected value of output conditional on inputs equals the value of the production function, that is,

$$E(y_i \,|\mathbf{x}_i) = E(f(\mathbf{x}_i)) + E(\varepsilon_i) = f(\mathbf{x}_i)\,.$$

Therefore, in this setting the production function f can be estimated by nonparametric regression techniques. Note that the term “regression” refers to the conditional mean $E(y_i \,|\mathbf{x}_i)$.

Hildreth (1954) was the first to consider nonparametric regression subject to monotonicity and concavity constraints in the case of a single input variable x (see also Hanson and Pledger, 1976). Kuosmanen (2008) extended Hidreth's approach to the multivariate setting with a vector-valued **x**, and coined the term *convex nonparametric least squares* (CNLS) for this method. CNLS builds upon the assumption that the true but unknown production function f belongs to the set of continuous, monotonic increasing and globally concave functions, F_2, imposing exactly the same production axioms as standard DEA.

The CNLS estimator of function f is obtained as the optimal solution to the infinite dimensional least squares problem

$$\begin{aligned}&\min_{f}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2\\ &\text{subject to}\\ &f \in F_2\end{aligned} \tag{2}$$

The functional form of f is not specified beforehand. Rather, the optimal solution will identify the best-fit function f from the family F_2. Note that set F_2 includes an infinite number of functions, which makes problem (2) impossible to solve through brute force trial and error. Further, problem (2) does not generally have a unique solution for any arbitrary input vector **x**, but a unique solution exists for estimating f for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$. Therefore, we will next discuss the estimation of f for the observed data points and extrapolation to unobserved points in sub-section 3.2.

3.1 CNLS estimator for the observed data points

A unique solution to problem (2) for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$, can be found by solving the following finite dimensional *quadratic programming* (QP) problem

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}}\sum_{i=1}^{n}(\varepsilon_i^{CNLS})^2\\ &\text{subject to}\\ &y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i^{CNLS}\ \ \forall i\\ &\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i\ \ \forall h,i\\ &\boldsymbol{\beta}_i \ge 0\ \forall i\end{aligned} \tag{3}$$

where α_i and $\boldsymbol{\beta}_i$ define the intercept and slope parameters of tangent hyperplanes that characterize the estimated piece-wise linear frontier (note that $\boldsymbol{\beta}_i'\mathbf{x}_i = \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + ... + \beta_{im}x_{im}$). Symbol ε_i^{CNLS} denotes the CNLS residual, which is an estimator of the true but unobserved $\varepsilon_i = v_i$. Note that in (3) the Greek letters are variables and the Latin letters are parameters (i.e., ($\mathbf{x}_i$, y_i) are observed data).

Kuosmanen (2008) introduced the QP formulation (3), and proved its equivalence with the infinite dimensional optimization problem (2). Specifically, if we denote the value of the objective function in the optimal solution to the infinite dimensional CNLS formulation (2) by SSE_{CNLS} (SSE = the sum of squares of errors), and that of the finite QP problem (3) by SSE_{QP}, then the equivalence can be stated as follows.

Theorem 1: $SSE_{CNLS} = SSE_{QP}$.

Proof. See Kuosmanen (2008), Theorem 2.1.

The equivalence result does not restrict to the objective functions, the optimal solution to problem (3) also provides us unique estimates of function *f* for the observed data points. Once the optimal solution is found, we will add "hats" on top of $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$, and refer to them as estimators.[7] In other words, α_i, $\boldsymbol{\beta}_i$, and ε_i^{CNLS} are variables of problem (3), whereas estimators $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$ provide the optimal solution to problem (3). Given $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ from (3), we define

$$\hat{f}^{CNLS}(\mathbf{x}_i) = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i'\mathbf{x}_i = y_i - \hat{\varepsilon}_i^{CNLS} . \tag{4}$$

This estimator of function *f* satisfies the following properties:

Theorem 2: *In the case of the neoclassical model with no inefficiency, $\hat{f}^{CNLS}(\mathbf{x}_i)$ is a unique, unbiased and consistent estimator of $f(\mathbf{x}_i)$ for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$.*

Proof. Uniqueness is proved by Lim and Glynn (2012), Proposition 1. Unbiasedness follows from Seijo and Sen (2011), Lemma 2.4. Consistency is proved under slightly different assumptions in Seijo and Sen (2011), Theorems 3.1 and 3.2, and Lim and Glynn (2012), Theorems 1 and 2.

The constraints of the QP problem (3) have the following compelling interpretations.[8] The first constraint of the least squares formulation (3) is a linear regression equation. However, the CNLS regression does not assume linear *f*: note that coefficients α_i and $\boldsymbol{\beta}_i$ are specific to each observation *i*. Using the terminology of DEA, α_i and $\boldsymbol{\beta}_i$ are directly analogous to the multiplier coefficients of the dual formulation of DEA. The inequality constraints in (3) can be interpreted as a system of *Afriat inequalities* (compare with Afriat, 1967, 1972; and Varian, 1984). As Kuosmanen (2008) emphasizes, the Afriat inequalities are the key to modeling the concavity axiom in the general multiple regression setting.

Coefficients α_i and $\boldsymbol{\beta}_i$ should not be misinterpreted as parameters of the estimated function *f*, but rather, as parameters characterizing tangent hyperplanes to an unknown production function *f*. These coefficients characterize a convex piece-wise linear function, to be examined in more detail the next sub-section. At this point, we must emphasize that we did not assume or restrict the domain F_2 to only include piece-wise linear function. In fact, it turns out that the "optimal" functional form to solving the infinite dimensional least squares problem (2) is always a convex piece-wise linear function characterized by coefficients α_i and $\boldsymbol{\beta}_i$. However, this optimal solution is unique only for the observed data points.

3.2 Extrapolating to unobserved points

In many applications we are interested in estimating the frontier not only for the observed data points, but also for unobserved input vectors **x**. Although the CNLS estimator is unique for the observed data points, there is no unique way of extrapolating the CNLS estimator to unobserved points. In general, the optimal solution to the infinite dimensional least squares problem (2) is not unique, but there exists a set of functions $f^* \in F_2^*$ that solve the optimization problem (2). Formally, we denote the set of alternate optima to (2) as

[7] In application, when estimators are calculated for a specific data set we will refer to these as estimated parameters.

[8] Note is formulation is written for ease of interpretation. Other formulations might be preferred to improve computational performance.

$$F_2^* = \left\{ f^* \middle| f^* = \arg\min_{f \in F_2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \right\}.$$

Kuosmanen (2008) characterizes the minimum and maximum bounds for the functions $f^* \in F_2^*$. It turns out that both bounds are piece-wise linear functions. However, only the minimum bound satisfies the postulated monotonicity and concavity properties. To resolve the non-uniqueness issue, Kuosmanen and Kortelainen (2012) appeal to the *minimum extrapolation principle* and propose to use the lower bound

$$\hat{f}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq \hat{f}^{CNLS}(\mathbf{x}_i) \ \forall i = 1,...,n \right\} \tag{5}$$

Note that the lower bound $\hat{f}_{\min}^{CNLS}$ is simply the DEA estimator (single output, variable returns to scale) applied to the observed inputs $\mathbf{x}_i$ and the fitted outputs $\hat{f}^{CNLS}(\mathbf{x}_i)$ obtained from equation (4).[9] The lower bound function satisfies the postulated properties of monotonicity and concavity. We can make the following connection between the lower bound (5) and the infinite dimensional CNLS problem (2).

Theorem 3: *Function* $\hat{f}_{\min}^{CNLS}$ *stated in equation* (5) *is one of the optimal solutions to the infinite dimensional optimization problem* (2). *It is the unique lower bound for the functions that solve problem* (2), *formally*

$$\hat{f}_{\min}^{CNLS}(\mathbf{x}) \leq f^*(\mathbf{x}) \text{ for all } \mathbf{x} \in \Re_+^m \text{ and } f^* \in F_2^*.$$

Proof. See Kuosmanen (2008) Theorem 4.1.

Note that while $\hat{f}^{CNLS}$ is unbiased and consistent for the observed points $\mathbf{x}_i$ (Theorem 3), the use of the piece-wise linear minimum function $\hat{f}_{\min}^{CNLS}$ will cause downward bias in finite samples as we apply the minimum extrapolation principle to extrapolate to unobserved points $\mathbf{x}$. Within the observed range of data, the downward bias will diminish as the sample size increases.

It is also worth noting that the optimal solution to the QP problem (3) does not necessarily produce unique coefficients $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$. Although $\hat{f}_{\min}^{CNLS}$ is a unique lower bound, consistent with the minimum extrapolation principle, the coefficients $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ obtained as the optimal solution to (5) need not be unique either. It is well-known in the DEA literature that these multiplier coefficients are not unique in the vertices of the piece-wise linear function.

3.4 Computational issues

The CNLS problem (3) has linear constraints and a quadratic objective function, hence it can be solved by QCP solvers such as CPLEX or MOSEK.[10] Standard solvers work well in relatively small sample sizes (50 – 200 firms) available in the majority of published applications of efficiency analysis. However, since the number of Afriat inequalities in (3) grows at a quadratic rate as a function of the number of observations, the computational burden becomes a significant issue when the sample size increases beyond 300 firms. Note that adding a new firm to the sample increases the

[9] In addition to the use of DEA to identify the lower bound function, there is a more fundamental connection between CNLS and DEA, to be explored in Section 4.

[10] Examples of computational codes for GAMS are available on the StoNED website: www.nomepre.net/stoned/.

number of unknown parameters by m+2, and the number of Afriat inequality constraints increases by 2n. Introducing an additional input variable increases the number of unknown parameters by n, but there is no impact on the number of constraints. For these reasons, standard QP algorithms are inadequate for handling large samples with several hundreds or thousands of observations.

As a first step towards improving computational performance in small samples and to allow for larger problems to be solved, Lee et al. (2013) propose to follow the strategy of Dantzig *et al.* (1954, 1959) to iteratively identify and add violated constraints. The algorithm developed by Lee et al. first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. In computational experiments, this algorithm allowed problems with up to 1,000 firms to be solved. Therefore, this algorithm has practical value especially in large sample applications and simulation-based methods such as bootstrapping or Monte Carlo studies. Another recent study by Hannah and Dunson (2013) implements CNLS in Matlab, reporting promising results. However, further algorithm development is needed to make the CNLS problem computable in very large sample sizes containing several thousands or millions of observations.

4. Deterministic frontiers

In this section we consider another special case of model (1) where the composite error term ε consists exclusively of inefficiency u, and there is no noise (i.e., v = 0). In the SFA literature, this special case is commonly referred to as the *deterministic model*. This does not imply, however, that probabilistic inferences are impossible.

Banker (1993) was the first to show that DEA can be understood as a maximum likelihood estimator of the deterministic model, with a statistical (probabilistic) foundation. However, the known statistical properties and inferences in the DEA literature restrict to the finite sample error that generally diminishes as the sample size increases. Or stated differently, the model specification and input and output data in the deterministic model are assumed to be exact and correct, so the only probabilistic component is the random sample of observations drawn from the production possibility set. This same deterministic model and its associated statistical foundation are used for inference in the bootstrapping methods (e.g., Simar and Wilson, 1998; 2000). Thus, statistical inference and confidence intervals estimated using bootstrapping methods only account for uncertainty in sampling and do not account for other sources of random variation or noise. Thus, bootstrap confidence intervals of DEA are not directly comparable to confidence intervals of other models that are genuinely stochastic in their nature (e.g., the SFA confidence intervals).

It is important to recognize that if the no noise assumption (v = 0) of the deterministic model does not hold, the statistical foundations of DEA collapse. The bootstrapping methods to adjust for the small sample are not a remedy against noise, rather adjusting for the sampling bias can make the DEA estimator worse if data are perturbed by noise. The stochastic case that includes both inefficiency and noise simultaneously will be considered in Section 5. The purpose of this section is to establish some useful connections between the ‘neoclassical’ CNLS and the ‘deterministic’ DEA to develop a unified framework and pave the way for a stochastic nonparametric StoNED estimator.

4.1 DEA as sign-constrained CNLS

In the single-output case, the variable returns to scale (VRS) DEA estimator of production function f can be stated as

$$\begin{aligned}\hat{f}^{DEA}(\mathbf{x}) &= \min_{\alpha,\boldsymbol{\beta}}\left\{\alpha+\boldsymbol{\beta}'\mathbf{x}\,\middle|\,\alpha+\boldsymbol{\beta}'\mathbf{x}_i \geq y_i \;\; \forall i=1,...,n\right\} \\ &= \max_{\boldsymbol{\lambda}}\left\{\sum_{h=1}^{n}\lambda_h y_h \,\middle|\, \mathbf{x} \geq \sum_{h=1}^{n}\lambda_h \mathbf{x}_h ; \sum_{h=1}^{n}\lambda_h = 1\right\}\end{aligned} \tag{6}$$

Note the difference between formulations (5) and (6): the former one uses the estimated output values $\hat{f}^{CNLS}(\mathbf{x}_i)$, whereas in the latter one uses the observed outputs y_i. Otherwise the formulations (5) and (6) are equivalent. The minimization formulation in (6) can be interpreted as the DEA multiplier formulation, whereas the maximization formulation of (6) is known as the DEA envelopment formulation. The duality theory of linear programming implies that the two formulations are equivalent.

Consider next a version of the CNLS estimator with an additional sign constraint on the residuals

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}}\sum_{i=1}^{n}(\varepsilon_i^{CNLS-})^2 \\ &\text{subject to} \\ &y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i^{CNLS-} \;\; \forall i \\ &\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\ &\boldsymbol{\beta}_i \geq 0 \;\forall i \\ &\varepsilon_i^{CNLS-} \leq 0 \;\forall i\end{aligned} \tag{7}$$

Comparing (3) and (6), we see that the only difference is the last constraint of (7), which is not present in the original CNLS formulation. Due to the sign constraint, Kuosmanen and Johnson (2010) interpret (6) as an axiomatic, nonparametric counterpart to the classic parametric programming approach of Aigner and Chu (1968).

We now establish the formal connection between CNLS and DEA as follows. Let $\hat{f}_{\min}^{CNLS-}(\mathbf{x})$ denote the piece-wise linear function obtained by applying equation (5) to the observed inputs $\mathbf{x}_i$ and the fitted values $\hat{y}_i$ of the sign-constrained formulation (7).

Theorem 4: *The sign-constrained CNLS estimator is equivalent to the DEA VRS estimator:*

$$\hat{f}_{\min}^{CNLS-}(\mathbf{x}) = \hat{f}^{DEA}(\mathbf{x})$$

Proof. Follows directly from Theorem 3.1 in Kuosmanen and Johnson (2010).

Although Theorem 4 was stated in the VRS case, the equivalence of DEA and sign-constrained CNLS does not restrict to the VRS case. Indeed parallel results are available for the other standard specifications of returns to scale by imposing additional constraints on the coefficients $\hat{\alpha}_i$ in formulations (3) or (7) as follows:

Constant returns to scale (CRS): impose $\hat{\alpha}_i = 0 \;\; \forall i$

Non-increasing returns to scale (NIRS): impose $\hat{\alpha}_i \geq 0 \;\; \forall i$

Non-decreasing returns to scale (NDRS): impose $\hat{\alpha}_i \leq 0 \;\; \forall i$

Similarly, if the convexity assumption of DEA is relaxed the free disposable hull (FDH), Afriat (1972), estimator provides the minimum envelopment of data subject to free disposability. Keshvari and Kuosmanen (2013) show that the FDH formulation is a sign-constrained special case of isotonic nonparametric least squares (INLS), which in turn is the concavity relaxed version of CNLS.

From a practical point of view, the least squares interpretation of DEA opens up new avenues for applying tools from econometrics to DEA. For example, Kuosmanen and Johnson (2010) propose to measure the goodness-of-fit of DEA estimator by using the standard *coefficient of determination* from regression analysis, specifically

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} . \tag{8}$$

Where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the average output in the sample. The R^2 statistic measures the proportion of output variation that is explained by the DEA frontier. While this variance decomposition can be applied to any regression model (including DEA), we note that DEA does not maximize the value of R^2 and hence negative R^2 values are possible for DEA estimators. This variance decomposition assumes a single output, however, one could compute and report separate R^2 statistics for each output.

4.2 Corrected CNLS

DEA builds on the minimum extrapolation principle to estimate the smallest function that envelops all data points. From the statistical point of view, insisting on the minimum extrapolation results in a systematic downward bias (i.e., the small sample error of DEA). For the deterministic model, Kuosmanen and Johnson (2010) show that a consistent and asymptotically unbiased estimator is obtained by applying a nonparametric variant of the classic COLS estimator. The proposed *corrected convex nonparametric least squares* (C^{2}NLS) estimator has always better discriminating power than DEA: the C^{2}NLS frontier envelops the DEA frontier everywhere, and the probability of finding multiple efficient units in randomly generated data approaches zero.

The C^{2}NLS method combines the nonparametric CNLS regression with the stepwise COLS approach first suggested by Winsten (1957), and more formally developed by Gabrielsen (1975) and Greene (1980). In this approach the most efficient firm in the sample is considered to be fully efficient, and the remaining inefficiency terms are normalized accordingly relative to the most efficient firm in the sample. A widely used panel data approach by Schmidt and Sickles (1984) applies a similar two-step approach (see Section 6.2 for details).

The essential steps of the C^{2}NLS routine can be described as follows:

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $E(y_i \,|\mathbf{x}_i)$.

Step 2: Identify the most efficient unit in the sample (i.e., $\hat{u}_{benchmark}^{C2NLS} = \max_{h \in \{1,\ldots,n\}} \hat{\varepsilon}_h^{CNLS}$) as the benchmark.

Adjust the CNLS residuals according to $\hat{u}_i^{C2NLS} = (\max_{h \in \{1,\ldots,n\}} \hat{\varepsilon}_h^{CNLS}) - \hat{\varepsilon}_i^{CNLS}$.

Step 3: Apply equation (5) to estimate the minimum function $\hat{f}_{\min}^{CNLS}(\mathbf{x})$. Adjust the minimum function by adding the residual of the benchmark firm to estimate the frontier using

$$\hat{f}^{C2NLS}(\mathbf{x}) = \hat{f}_{\min}^{CNLS}(\mathbf{x}) + \hat{u}_{benchmark}^{C2NLS}$$

Thus obtained $\hat{u}_i^{C2NLS}$ can be used as measures of inefficiency in the deterministic setting without noise. The most appealing properties of the C^{2}NLS estimator can be summarized as follows:

Theorem 5: *if* $\sigma_v = 0$, *then the C^{2}NLS estimator is statistically consistent:*

$$\operatorname*{plim}_{n\to\infty} \hat{f}^{\mathrm{C2NLS}}(\mathbf{x}_i) = f(\mathbf{x}_i) \text{ for all } i = 1,\ldots,n.$$

Proof. Follows from Theorem 4.1 in Kuosmanen and Johnson (2010).

Theorem 6: *the C^{2}NLS frontier envelops the DEA frontier, that is,*

$$\hat{f}^{C2NLS}(\mathbf{x}) \geq \hat{f}^{DEA}(\mathbf{x}) \; \forall \mathbf{x} \in \Re_+^m .$$

Proof. Follows from Theorem 4.2 in Kuosmanen and Johnson (2010).

Note that the inefficiency estimates $\hat{u}_i^{C2NLS}$ are non-negative by construction, with the value of zero indicating full efficiency. The inefficiency measures can be converted to Farrell (1957) output efficiency scores ($\hat{\theta}_i^{C2NLS} \in [0,1]$) by using

$$\hat{\theta}_i^{C2NLS} = \frac{y_i}{\hat{f}^{C2NLS}(\mathbf{x}_i)} = \frac{y_i}{y_i + \hat{u}_i^{C2NLS}} . \tag{9}$$

5. Stochastic Nonparametric Envelopment of Data (StoNED)

We are now equipped to consider the general stochastic nonparametric model that does not restrict to any particular functional form of *f* and includes both inefficiency *u* and stochastic noise *v*. Before proceeding to estimation, we must emphasize that the shift from the deterministic case to a stochastic model is rather dramatic. For example, measuring the distance from an observed point to the frontier does not provide a measure of inefficiency if the observed point is perturbed by noise. While probabilistic inference in the deterministic case only investigates finite sample error, in the stochastic model the noise term is still relevant even if the sample size approaches infinity. Clearly, when all data points are subject to noise enveloping all observations would overestimate the true frontier production function. The CNLS regression that fits a monotonic increasing and concave curve through the middle of the cloud of data provides a natural starting point for the next generation of DEA that can deal with noise.[11] Following Kuosmanen (2006), we refer to this approach as *stochastic nonparametric envelopment of data* (StoNED).

[11] Banker and Maindiratta (1992) consider maximum likelihood estimation of the unified frontier model subject to monotonicy and concavity constraints. However, their maximum likelihood problem appears to be computationally prohibitive. We are not aware of any application of this method. Gstach (1998) presents another early attempt to incorporate noise in DEA. However, he needs to make a rather restrictive assumption of truncated noise (see Simar and Wilson, 2011, for sharp critique of this assumption).

Analogous to the parametric COLS and MOLS (*modified OLS*) estimators and the nonparametric C^{2}NLS, the StoNED estimator consists of multiple steps. The main steps can be described as follows (a detailed description of each step follows below):

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $E(y_i \mid \mathbf{x}_i)$.

Step 2: Apply parametric methods (e.g., the method of moments or quasi-likelihood estimation) or nonparametric methods (e.g., kernel deconvolution) to the CNLS residuals ε_i^{CNLS} to estimate the expected value of inefficiency μ.

Step 3: Apply equation (5) to estimate the minimum function $\hat{g}_{\min}^{CNLS}(\mathbf{x})$. Adjust the minimum function by adding the expected inefficiency μ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\mu}$$

Step 4: Apply parametric methods (see e.g., Jondrow, Lovell, Materov and Schmidt, 1982, JLMS hereafter) or nonparametric deconvolution (e.g., kernel smoothing, Horrace and Parmeter, 2011) to estimate firm-specific inefficiency using the conditional mean $E(u_i \mid \varepsilon_i^{CNLS})$.

We will next describe each step in detail, noting that each step provides alternative modeling choices (depending on the assumptions one is willing to impose), and that it is not necessary to go through all of the steps. We discuss the information available at the end of each step and the possible motivations for proceeding to further steps.

5.1 Step 1: CNLS regression

The CNLS estimator was described in detail in Section 3 under the assumption of no inefficiency (*u* = 0). If the observed outputs are subject to asymmetric inefficiency, as the general frontier model (1) assumes, then the zero-mean assumption $E(\varepsilon_i) = 0$ of regression analysis is violated. Indeed, $E(\varepsilon_i) = E(v_i - u_i) = -E(u_i) < 0$ due to the asymmetric non-negative inefficiency term. Therefore, the CNLS estimator is no longer a consistent estimator of the frontier production function *f*.

Recall that CNLS regression estimates the conditional mean. Therefore, define the conditional mean function *g* as[12]

$$g(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i) = f(\mathbf{x}_i) - E(u_i). \tag{10}$$

If the random inefficiency term *u* is independent of inputs **x**, then the CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is an unbiased and consistent estimator of function *g*. The CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is obtained by solving the QP problem (3) and applying equation (4), as already discussed in Section 3, so we do not reproduce the CNLS formulations again here. Note that function *g* is simply the frontier production function *f* less the expected value of the inefficiency term *u*. If the inefficiency term *u* has a constant variance (i.e., inefficiency term *u* is homoscedastic), then the expected value of the inefficiency term *u* is a constant, denoted as μ. In other words, the CNLS provides a consistent estimator of the frontier *f* minus a constant. The constant μ can be estimated based on the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, as discussed in more detail in Section 5.2. The case of heteroscedastic inefficiency where $E(u_i)$ is no longer a constant will be examined in Section 8.

[12] Note that we use *g* to denote the conditional mean function when the composite error term contains inefficiency. This distinction was unnecessary in Section 3 because *g*(x) = *f*(x) when there is no inefficiency present.

Even if the data generating process (DGP) involves both inefficiency and noise, the CNLS estimator may be sufficient in some applications, without a need to proceed to the further stages. For example, if one is mainly interested in the relative efficiency rankings, then one could rank the evaluated units in descending order according to the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$. Further, if one is mainly interested in the marginal products of the input factors, the coefficients $\hat{\boldsymbol{\beta}}_i$ from (3), which are analogous to the multiplier coefficients (shadow prices) of DEA, then the CNLS regression provides consistent estimates (Seijo and Sen, 2011). The following steps described below do not influence the estimates of marginal products or the relative efficiency ranking of units. If one is interested in the frontier production function, average (in)efficiency in the sample, or cardinal firm-specific (in)efficiency estimates, then it is necessary to proceed further.

In the first step, one can impose some assumptions about returns to scale as described in Section 4.1. In addition, alternative modeling possibilities concern the multiplicative composite error and contextual variables are discussed as extensions in Section 6 and 7.

5.2 Step 2: Estimation of the expected inefficiency

Given the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, it is possible to estimate the expected value of the inefficiency term $\mu = E(u_i)$. Note that if the variance of the inefficiency is constant across firms (the homoscedasticity assumption), then the expectation is taken unconditional and is constant across firms.

Alternative approaches for estimating μ are available. We will next briefly review the commonly used parametric approaches based on the method of moments (Aigner et al., 1977), quasi-likelihood estimation (Fan et al., 1996), and the nonparametric kernel deconvolution (Hall and Simar, 2002).

5.2.1 Method of moments

The method of moments requires some additional parametric distributional assumptions. The moment conditions are known at least for the commonly used half-normal and exponential inefficiency distributions, but not for all distributions considered in the SFA literature (e.g., the gamma distribution). In the following, we will discuss the commonly assumed case of half-normal inefficiency and normal noise. Stated formally, we assume

$$u_i \sim N^+(0, \sigma_u^2)$$

and

$$v_i \sim N(0, \sigma_v^2)$$

The CNLS residuals are known to sum to zero $\sum_{i=1}^{n} \hat{\varepsilon}_i^{CNLS} = 0$ (Seijo and Sen, 2011). Hence, we can calculate the second and the third central moment of the residual distribution as

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^2 / (n-1) \tag{11}$$

$$\hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^3 / (n-1) \,. \tag{12}$$

The second central moment $\hat{M}_2$ is simply the sample variance of the residuals and the third central moment $\hat{M}_3$ is a component of the skewness measure. The hats on top of these statistics indicate these statistics are estimators of the true but unknown values of the central moments. If the parametric assumptions of half-normal inefficiency and normal noise hold, then the second and the third central moments are equal to

$$M_2 = \left[\frac{\pi-2}{\pi}\right]\sigma_u^2 + \sigma_v^2 \tag{13}$$

$$M_3 = \left(\sqrt{\frac{2}{\pi}}\right)\left[1-\frac{4}{\pi}\right]\sigma_u^3 \tag{14}$$

Note that the third moment only depends on the standard deviation of the inefficiency distribution (σ_u). Thus, given the estimated $\hat{M}_3$ (which should be negative), we can estimate σ_u as

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[1-\frac{4}{\pi}\right]}} \tag{15}$$

Subsequently, the standard deviation of the error term σ_v is estimated based on (12) as

$$\hat{\sigma}_v = \sqrt{\hat{M}_2 - \left[\frac{\pi-2}{\pi}\right]\hat{\sigma}_u^2}\ . \tag{16}$$

There has been considerable discussion in the recent literature regarding the question of how to proceed if $\hat{M}_3$ is positive. Carree (2002), Alminidis et al. (2009), and Alminidis and Sickles (2012) consider alternative inefficiency distributions that allow for positive skewness. Simar and Wilson (2010) maintain the standard distributional assumptions, but suggest instead the use of bootstrapping method.

5.2.2 Quasi-likelihood estimation

Another way to estimate the standard deviations σ_u, σ_v is to apply the quasi-likelihood method suggested by Fan et al. (1996) (who refer to it as pseudo-likelihood). In this approach we apply the standard maximum likelihood (ML) method to estimate the parameters σ_u, σ_v, taking the shape of the CNLS curve as given (thus the term quasi-likelihood, in contrast to the full information ML which would also parameterize the coefficients of the frontier).

One of the main contributions of Fan et al. (1996) was to show that the quasi-likelihood function can be stated as a function of a single parameter (i.e., the signal-to-noise ratio $\lambda = \sigma_u / \sigma_v$)[13] as,

$$\ln L(\lambda) = -n \ln \hat{\sigma} + \sum_{i=1}^{n} \ln \Phi\left[\frac{-\hat{\varepsilon}_i \lambda}{\hat{\sigma}}\right] - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\ , \tag{17}$$

where

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \left(\sqrt{2}\lambda\hat{\sigma}\right) \Big/ \left[\pi\left(1+\lambda^2\right)\right]^{1/2}, \tag{18}$$

$$\hat{\sigma} = \left\{ \frac{1}{n}\sum_{j=1}^{n}(\hat{\varepsilon}_i^{CNLS})^2 \Big/ \left[1-\frac{2\lambda^2}{\pi\left(1+\lambda\right)}\right]\right\}^{1/2}. \tag{19}$$

Symbol Φ denotes the cumulative distribution function of the standard normal distribution N(0,1). We first use (18) and (19) to substitute out $\hat{\varepsilon}_i$ and $\hat{\sigma}$ from (17). We then maximize the quasi-likelihood function (17) by enumerating over λ values, using a simple grid search or more

[13] The signal-to-noise ratio λ should not be confused with the intensity weights λ_i used in the envelopment formulation of DEA.

sophisticated search algorithms. When the quasi-likelihood estimate $\hat{\lambda}$ that maximizes (17) is found, we insert $\hat{\lambda}$ to equations (18) and (19) to obtain estimates of ε_i and σ. Subsequently, we can calculate estimates of $\hat{\sigma}_u = \hat{\sigma}\hat{\lambda}/(1+\hat{\lambda})$ and $\hat{\sigma}_v = \hat{\sigma}/(1+\hat{\lambda})$.

A simple practical trick to conduct quasi-likelihood estimation is to use ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). By specifying the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), we can trick the ML algorithm to perform the quasilikelihood estimation. This trick can also be used for estimating models involving contextual variables or heteroscedasticity (to be explored in Sections 7 and 8) by applying standard ML techniques as a second step.

5.2.3 Nonparametric kernel density estimation for the convoluted residual

While both method of moments and quasilikelihood techniques require parametric assumptions, a fully nonparametric alternative is available for estimating the signal-to-noise ratio λ , as proposed by Hall and Simar (2002). Their strategy is to search for a discontinuity in the residual density. The logic is that if an inefficiency term is left truncated, to represent efficient performance, there must be a discontinuity in distribution. When inefficiency is convoluted with noise, characterized by a continuous and smooth function, the discontinuity will still exist in the convoluted variable’s density, the estimated residuals density. Thus, Hall and Simar suggest estimating the density of the residual using kernel methods and use these estimates to identify the largest change in the derivative on the right-side of the distribution (in the case of a production function and left-side in the case of the cost function). Then under the assumption of homoscedastic noise and inefficiency, the location of the largest change in the derivative can be used to estimate the mean inefficiency in the sample.

More formally, note that residuals $\hat{\varepsilon}_i^{CNLS}$ are consistent estimators of $\varepsilon_i^+ = \varepsilon_i + \mu$. Thus, we can apply the kernel density estimator for estimating the density function of ε_i^+. Denote the kernel density estimator by f_{ε^+}. Hall and Simar (2002) show that the first derivative of the density function of the composite error term (f'_ε) is proportional to that of the inefficiency term (f'_u) in the neighborhood of μ . This is due to the assumption that f_u has a jump discontinuity at zero. Therefore, a robust nonparametric estimator of expected inefficiency μ is obtained as

$$\hat{\mu} = \arg\max_{z\in\Im}(\hat{f}'_{\varepsilon^+}(z)),$$

where $\Im$ is a closed interval in the right tail of f_{ε^+}.

5.3 Step 3: Estimating the frontier production function

In the presence of asymmetric inefficiency, the CNLS estimator estimates the conditional mean function $g(\mathbf{x}_i) = f(\mathbf{x}_i) - \mu$. Having estimated the expected inefficiency μ in Step 2, we can easily adjust the CNLS estimator to obtain an estimator of the frontier *f*. However, recall from Section 3 that the CNLS estimator of *g* is unique at the observed points $\mathbf{x}_i$ (*i*=1,…,*n*) but not in unobserved **x**. Therefore, Kuosmanen and Kortelainen (2012) recommend applying the lower bound of *g* (analogous to equation (5)), defined as

$$\hat{g}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}}\left\{\alpha + \boldsymbol{\beta}'\mathbf{x} \middle| \alpha + \boldsymbol{\beta}'\mathbf{x}_i \ge \hat{g}^{CNLS}(\mathbf{x}_i)\ \forall i = 1,...,n\right\}. \tag{20}$$

We can subsequently add the expected inefficiency μ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\mu} .$$

This equation summarizes the relation between the StoNED frontier and the CNLS estimator as well as the relation between the frontier function *f* and the conditional mean function *g*. The heteroscedastic case where the shapes of the frontier *f* and the regression $E(y_i | \mathbf{x}_i)$ are different will be discussed in Section 8 below.

5.4 Step 4: Estimating firm-specific inefficiencies

Measuring the distance from an observation to frontier is not enough for estimating efficiency in the stochastic setting because all observations are subject to noise. Hence the measured distance to frontier consists of both inefficiency and noise (plus any error in our frontier estimate).

We must emphasize that even though there exist statistically unbiased and consistent methods for the estimation of the frontier *f*, there is no consistent method for estimating firm-specific efficiencies *u* in the cross-sectional setting subject to noise. In a cross-section, estimating firm-specific realizations of a random variable u_i is impossible because we have only a single observation of each firm and all observations are perturbed by noise. This is not a fault of the methods (let alone their developers), it is just impossible to predict a realization of random variable based on a single observation that is subject to noise.

In the normal – half-normal case, Jondrow, Lovell, Materov and Schmidt (1982) (JLMS) develop a formula for the conditional distribution of inefficiency u_i given ε_i. The commonly used JLMS estimator for inefficiency is the conditional mean $E(u_i | \varepsilon_i)$. Given the parameter estimates $\hat{\sigma}_u$ and $\hat{\sigma}_v$, the conditional expected value of inefficiency can be calculated as[14]

$$E(u_i | \hat{\varepsilon}_i) = \frac{\hat{\sigma}_u \hat{\sigma}_v}{\sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \left[\frac{\phi\left(\dfrac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)}{1 - \Phi\left(\dfrac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)} - \frac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right], \tag{21}$$

where ϕ is the density function of the standard normal distribution *N*(0,1), Φ is the corresponding cumulative distribution function, and

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \hat{\sigma}_u \sqrt{2/\pi}$$

is the estimator of the composite error term (compare with (18)). It is worth to note that there is nothing "stochastic" in the equation (21): the JLMS formula is a simply a deterministic transformation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ to a new metric that represents the conditional expected value of the inefficiency term. Indeed, the rank correlation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ and the JLMS inefficiency estimates is equal to one (see Ondrich and Ruggiero, 2001). For the purposes of relative efficiency rankings, the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ are sufficient.

Horrace and Parmeter (2011) show that the parametric assumption of the inefficiency distribution can be relaxed. Their approach still requires the parametric assumption of normally distributed noise. Rather than assuming a specific parametric distribution for the inefficiency term, the authors assume the density of *u* belongs to the ordinary smooth family of distributions, which includes exponential, gamma or Laplace (see also Fan, 1991). They apply Hall and Simar's (2002)

[14] Note that equation (21) corrects the errors noted in formulations stated by Kuosmanen and Kortelainen (2012) and Keshvari and Kuosmanen (2013).

method to estimate the jump discontinuity and thus the signal to noise ratio. Given the mean inefficiency level the authors are then able to construct the full density distribution of the inefficiency term using kernel smoothing and the residuals from a conditional mean estimation.

5.5 Statistical specification tests of the frontier model

As discussed above, the StoNED estimator consists of four steps. If all firms are efficient and deviations from the frontier are due to noise, the step 1 of estimating the conditional mean function is sufficient, and there is no reason to proceed further to step 2 of estimating the mean inefficiency to step 3 shifting the conditional mean function or step 4 estimating firm specific inefficiencies. To determine whether one should proceed from step 1 further to step 2, the efficiency analyst may want to test the data for evidence of inefficiency. If the results of a statistical specification test indicate that there is significant inefficiency present, this can be a convincing argument even for skeptics who believe that markets function efficiently.

The residual $\hat{\varepsilon}_i^{CNLS}$ consists of two components, a normally distributed noise term and a left-truncated inefficiency term. Schmidt and Lin (1984) propose a test of the skewness of the residuals as a method to investigate if inefficiency is present. By only looking at the skewness, the method is robust to the common alternative specifications of the inefficiency term in the stochastic frontier model. Thus, the null hypothesis is the residuals are normally distributed and a $\sqrt{b_1}$ test calculated as

$$\sqrt{b_1} = \frac{m_3}{(m_2)^{3/2}} \tag{22}$$

Where m_2 and m_3 are, the second and third moments of the residuals respectively. The distribution of the skewness test statistic, $\sqrt{b_1}$ can be constructed by a simple Monte Carlo simulation as described in D'Agostino and Pearson (1973). The authors also provide tables with critical values of the proposed test statistic for different sample sizes.

Kuosmanen and Fosgerau (2009) consider a fully nonparametric specification test that relaxes the normality assumption of the noise term. They show that the same test statistic $\sqrt{b_1}$ considered by Schmidt and Lin (1984) can be used for testing the null hypothesis of a symmetric v against the alternative hypothesis of skewness. They also recognize the $\sqrt{b_1}$ can wrongly reject the null hypothesis if the distribution is symmetric but has fat tails. Thus, they propose the additional b_2 test of the fourth moment

$$b_2 = \frac{m_4}{(m_2)^2} \tag{23}$$

Where m_2 and m_4 are the second and fourth moments of the residuals respectively. The null hypothesis is that the distribution is normally distributed. The alternative hypothesis is that there is non-normal kurtosis. The results of the $\sqrt{b_1}$ and b_2 tests can be given the following interpretation:

- If the null hypothesis of normality is rejected in the $\sqrt{b_1}$ test but maintained in the b_2 test, there is strong evidence in favor of a frontier model.
- If the null hypothesis of normality is maintained both in the $\sqrt{b_1}$ and b_2 tests, this supports the hypothesis of a competitive market with no inefficiency present.
- If the null hypothesis is rejected in the b_2 test, there may be data problems or model misspecification. There is no conclusive evidence in favor or against the frontier model.

It is worth noting that the power of the test depends on how specifically the null hypothesis and the alternative hypothesis are stated. For example, the $\sqrt{b_1}$ test of normality is more powerful than the fully nonparametric test of symmetry. If we are willing to impose some distributional assumptions for the inefficiency term, then more powerful specification tests are available. For example, Coelli (1995) proposed a variant of the Wald test to test the null hypothesis that there is no inefficiency, i.e. $\sigma_u^2 = 0$, against the alternative $\sigma_u^2 > 0$. While imposing distributional assumptions can increase the power of the test, it will also increase the risk of misspecification, which would make the statistical test inconsistent.

6. Extensions

6.1 Multiplicative composite error term

Most SFA studies use Cobb-Douglas or translog functional forms where inefficiency and noise affect production in a multiplicative fashion. In the present context, it is worth noting that the assumption of constant returns to scale (CRS) would also require multiplicative error structure, as will be discussed in more detail below. Further, a multiplicative error specification implies a specific model of heteroscedasticity in which the variance of the composite error term increases with firm size.

Multiplicative composite error structure is obtained by rephrasing model (1) as

$$y_i = f(\mathbf{x}_i) \cdot \exp(\varepsilon_i) = f(\mathbf{x}_i) \cdot \exp(v_i - u_i) \tag{24}$$

Applying the log-transformation to equation (23), we obtain

$$\ln y_i = \ln f(\mathbf{x}_i) + \varepsilon_i . \tag{25}$$

Note that the log-transformation cannot be applied directly to inputs **x** – it must be applied to the production function *f*.

In the multiplicative case, the CNLS formulation (3) can be rephrased as

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \varepsilon_i^{CNLS} \;\; \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h, i \\
&\boldsymbol{\beta}_i \ge 0 \;\; \forall i
\end{aligned} \tag{26}$$

where $\phi_i + 1$ is the CNLS estimator of $E(y_i \,|\mathbf{x}_i)$. The value of one is added here to make sure that the computational algorithms do not try to take logarithm of zero. The first equality can be interpreted as the log transformed regression equation (using the natural logarithm function ln(.)). The second through fifth constraints are similar to (3) with the exception observed output in (3) is replaced with $\phi_i + 1$. The use of ϕ_i allows the estimation of a multiplicative relationship between output and input while assuring convexity of the production possibility set in original input-output space.[15]

Note that the log-transformation of a model variable renders the optimization formulation as a nonlinear programming (NLP) problem. These constraints are shown separately to illustrate the

[15] If we apply the log transformation directly to input data, the resulting frontier would be a piece-wise log-linear frontier, which has been considered in the DEA literature by Charnes et al. (1982) and Banker and Maindiratta (1986). Unfortunately, the piece-wise log-linear frontier does not generally satisfy the concavity of *f*.

connection to previous formulations, but the first equality constraint can be moved to the objective function by solving and substituting for $\hat{\varepsilon}_i^{CNLS}$. Thus we have a convex solution space and a nonlinear objective function. This formulation can be solved by standard nonlinear programming algorithms and solvers. NLP solvers are available for example in such mathematical programming packages as GAMS, AIMMS, Matlab, and Lindo, among others.

In the multiplicative case, the CNLS estimator (25) can be applied, or as the first step of the C^2NLS or StoNED estimation routine. The standard method of moment, quasi-likelihood and kernel deconvolution techniques apply, as described in Section 5. However, note that in step 3 the frontier production function is obtained as $\hat{f}^{StoNED}(\mathbf{x}_i) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) \cdot \exp(\hat{\mu})$, where $\hat{g}_{\min}^{CNLS}(\mathbf{x})$ is the minimum function computed using equation (19.5) and $\exp(\hat{\mu})$ is the estimated average efficiency. A convenient feature of the multiplicative model is that exp(u_i) can be interpreted as the Farrell output efficiency measure.

6.2 Panel data

In panel data the sample of firms is observed repeatedly over multiple time periods. Panel data applications are common in the SFA literature and a number of alternative SFA models involving time invariant and time varying inefficiency are available (see, e.g., Greene, 2008, Section 2.7). In contrast, DEA studies ignore the time dimension of the panel data and either pool the panel together as a single cross section or treat each time period as an independent cross section.[16]

The regression interpretation of DEA examined in Section 4.1 allows us to combine DEA-style axiomatic frontier with the modern panel data methods from econometrics. Kuosmanen and Kortelainen (2012, Section 4.1) were the first consider a fixed effects approach to estimating a time invariant inefficiency model. Their fully nonparametric panel data StoNED estimator can be seen as a nonparametric counterpart to the classic SFA approach by Schmidt and Sickles (1984). In the following we consider the random effects approach, building upon Eskelinen and Kuosmanen (2013).

Consider a data set where each firm is observed over time periods $t = 1,...,T$ and define a time invariant frontier model

$$y_{it} = f(\mathbf{x}_{it}) - u_i + v_{it} \ \forall i = 1,...,n \ \forall t = 1,...,T, \tag{27}$$

where y_{it} is the observed output of firm *i* in time period *t*, $\mathbf{x}_{it}$ is a vector of inputs consumed by firm *i* in time period *t*, and *f* is a frontier production function that is time invariant and common to all firms. As before, u_i is a firm specific inefficiency term that does not change over time, and v_{it} is a random disturbance term of firm *i* in period *t*. Similar to the cross-sectional model, we assume that u_i and v_{it} are independent of inputs $\mathbf{x}_{it}$ and of each other.[17]

To estimate the model (27), we can adapt the standard CNLS estimator as

[16] One notable exception is Ruggiero (2004).

[17] The random effects approach to panel data requires that the time invariant inefficiency is uncorrelated with inputs. This is a strong assumption. Marschak and Andrews (1944) were among the first to note that rational firm manager will adjust the inputs to take into account the technical inefficiency, and hence the observed inputs are correlated with inefficiency. In that case, the random effects estimator is biased and inconsistent. The fixed effects estimator considered by Kuosmanen and Kortelainen (2012) does not depend on this assumption.

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{t=1}^{T}\sum_{i=1}^{n}(\varepsilon_{it}^{CNLS})^2$$

subject to

$$y_{it} = \alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} + \varepsilon_{it}^{CNLS} \;\; \forall i = 1,...,n \;\; \forall t = 1,...,T$$
$$\alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} \le \alpha_{it} + \boldsymbol{\beta}_{it}'\mathbf{x}_{hs} \;\; \forall h,i = 1,...,n \;\; \forall s,t = 1,...,T \tag{28}$$
$$\boldsymbol{\beta}_{it} \ge \mathbf{0} \;\; \forall i = 1,...,n \;\; \forall t = 1,...,T$$

where $\hat{\varepsilon}_{it}^{CNLS}$ is the CNLS residual of firm *i* in period *t*. Note the parameters α_{it} and $\boldsymbol{\beta}_{it}$ that define the tangent hyperplanes of the estimated production function are specific to each firm in each time period. Thus, a piece-wise linear frontier is estimated with as many as *nT* hyperplanes.

Given the optimal solution to (28), we compute the firm-specific effects as

$$\bar{\varepsilon}_{i}^{CNLS} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^{CNLS} \tag{29}$$

Following Schmidt and Sickles (1984) we measure efficiency relative to the most efficient firm in the sample (analogous to the C^{2}NLS approach considered in Section 4.2) and define

$$\hat{u}_{i}^{StoNED} = (\max_{h\in\{1,...,n\}} \bar{\varepsilon}_{h}^{CNLS}) - \bar{\varepsilon}_{i}^{CNLS}. \tag{30}$$

To estimate theconditional mean function, we can adapt equation (20) to panel data as

$$\hat{g}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}}\left\{\alpha + \boldsymbol{\beta}'\mathbf{x} \,\middle|\, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} \ge \hat{g}^{CNLS}(\mathbf{x}_{it}) \;\forall i = 1,...,n; \forall t = 1,...,T\right\}.$$

The StoNED frontier estimator is then obtained as

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + (\max_{h\in\{1,...,n\}} \bar{\varepsilon}_{h}^{CNLS}).$$

Both the frontier and inefficiency estimators can be shown to be statistically consistent under the assumptions stated above.

Note that the panel data StoNED estimator described above is fully nonparametric in the sense that no parametric functional form or distributional assumptions are required. Still, the model described in equation (27) relies on two strong assumptions: i) there is no technical progress, and ii) inefficiency is constant over time. It is possible to relax these assumptions, but this will require some additional assumptions (typically imposing some parametric structure). Note that random effects estimator considered above may still be useful even if inefficiency changes over time. In that case, the inefficiency estimator can be interpreted as the average efficiency during the time period under study. Eskelinen and Kuosmanen (2013) propose to examine the development trajectories of the normalized CNLS residuals $\hat{\varepsilon}_{it}^{CNLS} / (\max_{h\in\{1,...,n\}} \bar{\varepsilon}_{h}^{CNLS})$ to gain a better understanding how the firm performance has developed during the study period. While the normalized CNLS residuals contain random noise, a growth trend (or decline) provides a clear indication that the performance of the firm has improved (or deteriorated) during the study period.

Based on the previous discussion, two insights are worth noting:

1) Panel data is not a panacea: while we recognize that panel data provides a richer set of information, we must also acknowledge that the intertemporal setting involves complex dynamics such as technological progress and changes in efficiency over time. The random effects approach to panel data considered above would be ideal for modeling experimental data where the researcher can control the input levels and keep the production technology the same across repeated experiments. However, most panel data applications of stochastic frontiers use observational data where both the production function and the level of efficiency will likely change over time.

2) Resorting to a fully nonparametric approach does not imply freedom from restrictive assumptions. In fact, avoidance of parametric assumptions often comes at the cost of very restrictive assumptions of no noise, no technical progress, or time invariant inefficiency. Indeed, insisting on a fully nonparametric approach can be more restrictive than resorting to some parametric assumptions that allow for explicit modeling of noise, technical progress, or time varying inefficiency.

6.3 Multiple outputs (DDF formulation)

The ability to model multiple inputs and multiple outputs has long been touted as an advantage of DEA over SFA: several DEA papers erroneously state that SFA cannot deal with multiple outputs. Lovell et al (1994) and Coelli and Perelman (1999; 2000) were the first to consider a stochastic distance function model that characterizes a general multiple inputs and multiple outputs technology using the radial input and output distance functions. The recent paper by Kuosmanen, Johnson and Parmeter (2013) (henceforth KJP) examines the assumptions of the data generation process that need to be satisfied for econometric identification of the distance function when the data are subject to random noise. Although the econometric estimation of distance functions is feasible, the well-established drawbacks of SFA still apply: a functional form needs to be specified for the distance function and parametric assumptions are typically made to decompose the residual into inefficiency and noise. Further, the commonly used parametric functional forms have the wrong curvature in output space, which is a serious problem for modeling joint production of multiple outputs.[18]

Up to this point, the CNLS/StoNED framework has been presented in the single output, multiple input setting. In this section we describe the CNLS estimator within the directional distance function (DDF) framework, Chambers et al. (1996, 1998). The CNLS formulation satisfies the axiomatic properties of the DDF by construction, models multiple inputs and multiple outputs, and accounts for stochastic noise explicitly, addressing the key limitations of both DEA and the parametric approaches. In the following we will briefly describe the stochastic data generating process (DGP) and the estimation of the DDF by CNLS. See KJP for a more detailed discussion.

The DDF indicates the distance from a given input-output vector to the boundary of the production possibility set T in some pre-assigned direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$, formally,

$$\vec{D}_T(\mathbf{x}, \mathbf{y}, \mathbf{g}^x, \mathbf{g}^y) = \sup_\theta \left\{ \theta \,\middle|\, (\mathbf{x} - \theta \mathbf{g}^x, \mathbf{y} + \theta \mathbf{g}^y) \in T \right\}. \tag{31}$$

Denote the reference input-output vector of firm i in the direction $(\mathbf{g}^x, \mathbf{g}^y)$ by $(\mathbf{x}_i^*, \mathbf{y}_i^*)$. In this section we do not impose any particular behavioral hypothesis, but it may be illustrative to interpret $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ as the optimal solution to firm i's profit maximization problem. Regardless of the firm manager's objective, we assume $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ lies on the boundary of the production possibility set T and hence the values of the DDF satisfy

$$\vec{D}_T(\mathbf{x}_i^*, \mathbf{y}_i^*, \mathbf{g}^x, \mathbf{g}^y) = 0 \quad \forall i = 1, \ldots, n \tag{32}$$

[18] The wrong curvature violates some of the most elementary properties of production technologies. For example, the Cobb-Douglas or translog specifications of the distance function will violates the basic properties of null jointness and unboundedness (see, e.g., Färe et al., 2005). Another problem concerns the economies of scope (e.g., Panzar and Willig, 1981). For example, the Cobb-Douglas distance function cannot capture the economies of scope at any parameter values. Since the economic rationale for joint production is rooted to economies of scope, it is contradictory to apply a technology that exhibits economies of specialization for modeling joint production.

The observed input-output vectors $(\mathbf{x}_i, \mathbf{y}_i),\ i = 1,...,n$, are perturbed in direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$ by random inefficiency u_i and noise v_i, which form the composite error term $\varepsilon_i = u_i + v_i$ (note the positive sign of the inefficiency term u_i). Specifically, the observed data are perturbed versions of the optimal input-output vectors as follows

$$(\mathbf{x}_i, \mathbf{y}_i) = (\mathbf{x}_i^* + \varepsilon_i \mathbf{g}^x, \mathbf{y}_i^* - \varepsilon_i \mathbf{g}^y) \quad \forall i = 1, \dots, n \tag{33}$$

We assume the inefficiency and noise terms satisfy the assumptions discussed in Section 2. Note that the elements of the direction vector $(\mathbf{g}^x, \mathbf{g}^y)$ represent the impacts of inefficiency and noise on specific input and output variables. If an element of $(\mathbf{g}^x, \mathbf{g}^y)$ is equal to zero, it means that the corresponding input or output variable is immune to both inefficiency and noise in the DGP. The larger the value of an element of $(\mathbf{g}^x, \mathbf{g}^y)$ in the DGP, the larger the impact of inefficiency and noise on the corresponding input or output variable is. Interestingly, Proposition 3 in KJP shows that in the DGP described above the value of the DDF equals the composite error term:

$$\vec{D}_T(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) = \varepsilon_i \quad \forall i\,.$$

This result provides implicitly a regression equation for estimating the DDF. We can resort to a similar stepwise procedure as described in Section 5.

The first step is to estimate the conditional mean distance defined as

$$d(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) = \vec{D}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) - \mu \tag{34}$$

Let Δ denote the set of functions that satisfy the axioms of free disposability, convexity, and the translation property.[19] We can adapt the CNLS estimator to the DDF setting by postulating the following infinite dimensional least squares problem

$$\begin{aligned} &\min_d \sum_{i=1}^n d(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y)^2 \\ &\text{subject to} \\ &d \in \Delta \end{aligned} \tag{35}$$

Formulation (35) is a complex, infinite dimensional optimization problem that cannot be solved by brute-force numerical methods. The main challenge is to find a way to parameterize the infinitely large set of functions that satisfy the stated regularity conditions. Here again we apply insights from Kuosmanen (2008) and show an equivalent finite dimensional representation in terms of quadratic programming. Consider the following QP problem

[19] The translation property, Chambers et al. (1998), states that if we move from the initial point ($\mathbf{x}$,$\mathbf{y}$) in the direction ($\mathbf{g}^x$, $\mathbf{g}^y$) by factor α , i.e., to the point $(\mathbf{x} + \alpha \mathbf{g}^x, \mathbf{y} - \alpha \mathbf{g}^y)$, then the distance to the frontier decreases by α . This property is crucial for the internal consistency of the DDF and can be seen as an additive analogue of the linear homogeneity property of the input distance function.

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\boldsymbol{\gamma}_i'\mathbf{y}_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i - \varepsilon_i^{CNLS} \;\; \forall i = 1,\dots,n \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i - \boldsymbol{\gamma}_i'\mathbf{y} \le \alpha_h + \boldsymbol{\beta}_i'\mathbf{x}_i - \boldsymbol{\gamma}_h'\mathbf{y}_i \;\; \forall h,i = 1,\dots,n \\
&\boldsymbol{\gamma}_i'\mathbf{g}^y + \boldsymbol{\beta}_i'\mathbf{g}^x = 1 \;\forall i = 1,\dots,n \\
&\boldsymbol{\beta}_i \ge \mathbf{0} \;\forall i = 1,\dots,n \\
&\boldsymbol{\gamma}_i \ge \mathbf{0} \;\forall i = 1,\dots,n
\end{aligned} \tag{36}$$

Note that the residual $\hat{\varepsilon}_i^{CNLS}$ here represents the estimated value of d_i (i.e., $\vec{D}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) + u_i$). We also introduce new firm-specific coefficients $\boldsymbol{\gamma}_i$ that represent marginal effects of outputs to the DDF. The first constraint defines the distance to the frontier as a linear function of inputs and outputs. The linear approximation of the frontier is based on the tangent hyperplanes, analogous to the original CNLS formulation. The second set of constraints is the system of Afriat inequalities that impose global concavity. The third constraint is a normalization constraint that ensures the translation property. The last two constraints impose monotonicity in all inputs and outputs. It is straightforward to show that the CNLS estimator of function d satisfies the axioms of free disposability, convexity, and the translation property (see Theorem 3 in KJP).

After solving the CNLS problem, one can proceed to estimate the deterministic frontier by Corrected CNLS as described in Section 4.2 or the stochastic frontier by StoNED as described in Section 5.2. Note that the CNLS estimator described above does not estimate the DDF directly, but rather $\vec{D}\,(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) + E(u_i)$. If the inefficiency term is homoscedastic, then the techniques described in Section 5.2 apply for the estimation of $E(u_i) = \mu$. The case of heteroskedastic inefficiency term is discussed in Sections 8.2 and 8.3 below. Subsequently, the estimate of the DDF is obtained by shifting the CNLS estimate of function d in direction $(\mathbf{g}^x, \mathbf{g}^y)$ by the estimated expected inefficiency.

To connect the multi-output DDF to the single output case, it is worth noting in the single output case, specifying the direction vector as g^y=1 and $\mathbf{g}^x$=$\mathbf{0}$, the CNLS problem (36) reduces to

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i - \varepsilon_i^{CNLS} \;\; \forall i = 1,\dots,n \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i = 1,\dots,n \\
&\boldsymbol{\beta}_i \ge \mathbf{0} \;\forall i = 1,\dots,n
\end{aligned} \tag{37}$$

This formulation is equivalent to the CNLS formulation (3) developed in Kuosmanen (2008), except for the sign of the residual $\hat{\varepsilon}_i^{CNLS}$ in the first constraint. Note that the DDF has positive values below the frontier and negative values above the frontier, which explains the negative sign.

6.4 Convex nonparametric quantile regression and asymmetric least squares

While CNLS estimates the conditional mean $E(y_i \,|\mathbf{x}_i)$, quantile regression aims at estimating the conditional median or other quantiles of the response variable (Koenker and Bassett, 1978;

Koenker, 2005).[20] Denoting the pre-assigned quantile by parameter $q \in (0,1)$, we can modify the CNLS problem (3) to estimate convex nonparametric quantile regression (CNQR) (Wang et al., 2014) as follows:[21]

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} q\sum_{i=1}^{n}\varepsilon_i^+ + (1-q)\sum_{i=1}^{n}\varepsilon_i^- \\
&\text{subject to} \\
&y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i^+ - \varepsilon_i^- \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \;\; \forall i \\
&\varepsilon_i^+ \ge 0 \;\; \forall i \\
&\varepsilon_i^- \ge 0 \;\; \forall i
\end{aligned} \tag{38}$$

The CNQR problem differs from CNLS in that the composite error term is now broken down to two non-negative components $\varepsilon_i^+, \varepsilon_i^- \ge 0$. The objective function minimizes the asymmetric absolute deviations from the frontier instead of symmetric quadratic deviations. The pre-assigned weight q defines the quantile to be estimated. For example, by setting q = 0.05, the piece-wise linear CNQR function will allow at most 5 percent of observations to lie above the fitted function and envelope at most 95 percent of the observed data points. As the sample size approaches to infinity, the q-order frontier will envelop exactly q percent of the observed data points (Wang et al., 2014, Theorem 1). Two important special cases are worth noting. First, if we set q = 0.5, then CNQR estimates the conditional median (whereas CNLS estimates the conditional mean). Secondly, as q approaches to zero, the negative deviations ε_i^- get a larger weight, and the CNQR approaches to the DEA frontier.

An appealing feature of the CNQR formulation is that its objective function and all constraints are linear functions of unknown parameters, and hence the CNQR problem can be solved by standard linear programming (LP) algorithms. However, a major drawback compared to CNLS is that the optimal solution to the CNQR problem is not necessarily unique, not even for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$. In econometrics, non-uniqueness of quantile regression is usually assumed away by assuming the regressors **x** are randomly drawn from a continuous distribution. In practice, however, input vectors **x** are not randomly drawn, and there may be two or more firms use exactly the same amounts of inputs (i.e., $\mathbf{x}_i = \mathbf{x}_j$ for firms i and j). In our experience, non-uniqueness of CNQR seems to be particularly a problem in samples where inputs **x** are discrete variables. Wang et al. (2014) recognize non-uniqueness of the CNQR estimator, illustrating the problem with a numerical example.

One possible way to resolve the non-uniqueness problem is to apply the asymmetric least squares criterion suggested by Newey and Powell (1987), and reformulate the CNQR problem as

[20] In the DEA literature, the quantile frontiers are commonly referred to as robust order-*m* and order-α frontiers (e.g., Aragon et al. 2005; Daouia and Simar, 2007). However, while quantile frontiers are more robust to outliers than the conventional DEA frontiers, the quantile DEA approaches typically assume away noise.

[21] Similar quantile formulation was first considered by Banker et al. (1991), who refer to it as ”stochastic DEA”.

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} q\sum_{i=1}^{n}(\varepsilon_i^+)^2 + (1-q)\sum_{i=1}^{n}(\varepsilon_i^-)^2 \\
&\text{subject to} \\
&y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i^+ - \varepsilon_i^- \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \;\; \forall i \\
&\varepsilon_i^+ \ge 0 \;\; \forall i \\
&\varepsilon_i^- \ge 0 \;\; \forall i
\end{aligned} \tag{39}$$

To our knowledge, this asymmetric least squares formulation has not been considered before; we will henceforth refer to it as convex asymmetrically weighted least squares (CAWLS). The CAWLS problem differs from CNQR only in terms of the objective function, which now minimizes the asymmetric squared deviation instead of the absolute deviations. In the case of the linear regression, Newey and Powell (1987) show that the properties of the asymmetric least squares estimator are analogous to those of the quantile regression, but the asymmetric least squares can be more convenient for statistical inferences. In the present context, we hypothesize that the use of the quadratic loss function similar to CNLS ensures that the optimal solution to the CAWLS problem is always unique for the observed data points $(\mathbf{x}_i, y_i),\ i = 1,...,n$. We leave confirming or rejecting this hypothesis as an open question for future research. Besides the question of uniqueness, the statistical properties of both CNQR and CAWLS would require further research.

CNQR and CAWLS formulations allow one to estimate the *q*-quantile or *q*-expectile frontiers directly, without a need to impose parametric distributional assumptions for the inefficiency and noise terms or resort to stepwise estimation along the lines described in Section 5. This is one of the attractive properties of CNQR and CAWLS. For the purposes of efficiency analysis, however, the use of quantiles or asymmetric weighted least squares is not a panacea. It is important to stress that the distance from the frontier, measured as $\hat{\varepsilon}_i^{CNQR} = \hat{\varepsilon}_i^+ - \hat{\varepsilon}_i^-$ or $\hat{\varepsilon}_i^{CAWLS} = \hat{\varepsilon}_i^+ - \hat{\varepsilon}_i^-$ (note: in both cases the residuals satisfy $\hat{\varepsilon}_i^+\hat{\varepsilon}_i^- = 0\ \forall i$), should not be interpreted as a measure of inefficiency, as the distance to frontier also includes noise. To estimate conditional expected value of inefficiency along the lines of JLMS, we still need to resort to stepwise estimation. One possibility is to replace CNLS by CNQR or CAWLS as the first step of the StoNED procedure outlined in Section 5. Of course, residuals $\hat{\varepsilon}_i^{CNQR}$ or $\hat{\varepsilon}_i^{CAWLS}$ can be used as such for relative performance rankings, but such performance rankings obviously depend on the chosen parameter value of *q*. Wang et al. (2014) examine the specification of *q* for frontier estimation, showing that the optimal value of *q* is a monotonically decreasing function of the signal to noise ratio $\lambda = \sigma_u / \sigma_v$. One may set the value of *q* based on subjective judgment, but in real world applications (consider, e.g., regulation of electricity distribution networks; see Kuosmanen, 2012; Kuosmanen, Saastamoinen and Sipiläinen, 2013), some objective criteria for specifying *q* would be important.

One appealing feature of the *q*-quantile and *q*-expectile frontiers is that they are robust to heteroscedasticity. Therefore, testing of and dealing with heteroscedasticity provide one promising application area for the CNQR and CAWLS techniques. If the composite error term is homoscedastic, then the quantile and expectile frontiers should have similar shapes at different values of *q*. Newey and Powell (1987) apply this idea for testing heteroscedasticity. We return to this issue in more detail in Section 8.

7. Contextual variables

A firm's ability to operate efficiently often depends on operational conditions and practices, such as the production environment and the firm specific characteristics for example technology selection or managerial practices. Banker and Natarajan (2008) refer to both variables that characterize operational conditions and practices as *contextual variables*. Currently two-stage DEA (2-DEA) is widely applied to investigate the importance of contextual variables as summarized by the citations included in Simar and Wilson (2007). However, its statistical foundation has been subject to sharp debate between Simar and Wilson (2007, 2011) and Banker and Natarajan (2008) (see also Hoff, 2007; McDonald, 2009). In this section we shed some new light on this debate following Johnson and Kuosmanen (2011, 2012).

It is important to note that Simar and Wilson (2007, 2011) do not consider stochastic noise in their DGP. In contrast, Banker and Natarajan (2008) introduce a noise term that has a doubly-truncated distribution, following the DEA+ approach by Gstach (1998). In this setting, Johnson and Kuosmanen (2012) show that the 2-DEA estimator of contextual variables is consistent under more general assumption that those stated by Banker and Natarajan (2008) and criticized by Simar and Wilson (2011). Further, Johnson and Kuosmanen (2012) employ the least squares formulation of DEA to develop a one-stage DEA method (1-DEA) for estimating the effects of the contextual variables. Relaxing the peculiar assumption of truncated noise,[22] Johnson and Kuosmanen (2011) develop *stochastic (semi-) nonparametric envelopment of z-variables data* (StoNEZD).

Taking the multiplicative model described in Section 6.1 as our starting point, we introduce the contextual variables, represented by *r*-dimensional vectors $\mathbf{z}_i$ that represent the measured values of operational conditions and practices, to obtain the following semi-nonparametric, partial log-linear equation

$$\ln y_i = \ln f(\mathbf{x}_i) + \boldsymbol{\delta}'\mathbf{z}_i + v_i - u_i . \qquad (40)$$

In this equation, parameter vector $\boldsymbol{\delta} = (\delta_1 \ldots \delta_r)'$ represents the marginal effects of contextual variables on output. All other variables maintain their previous definitions.

In the following sub-sections we will present two-stage DEA (2-DEA), one-stage DEA, and StoNEZD estimators. First, the 2-DEA estimator is described and the statistical properties of it are discussed. Given the assumptions necessary for the consistency of two-stage DEA method we then present the one-stage alternative. The joint estimation avoids the bias in the DEA frontier being transmitted to the parameter estimates of the coefficients on the contextual variables; however, the frontier estimated is still the minimum envelopment of the data and thus does not account for noise in the production model or input/output data. To account for stochastic noise, StoNEZD is introduced in 7.3.

7.1 Two-stage DEA

The literature on 2-DEA includes a number of variants. This sub-section follows the approach by Banker and Natarajan (2008). The two stages of their 2-DEA method are the following. In the first

[22] We label this assumption as peculiar because it contradicts standard statistical assumptions, namely, the residual term is often model as normally distributed because a mixture of a large number of unknown distributions is approximately normal in finite samples and asymptotically normal. The large number of unknown distributions is a result of measurement errors, modeling simplifications, and other sources of noise. Thus, the motivation for truncated normal distribution used in Gstach (1998) and Banker and Natarajan (2008) is lacking and peculiar as also noted by Simar and Wilson (2011). Johnson and Kuosmanen (2012) argue this truncation may come from an outlier detection procedure that would remove extreme observations from the analysis. However, in this case 1-DEA (introduced below) would still be preferred to 2-DEA because the bias introduced in two-stage estimation.

stage, the frontier production function f is estimated using the nonparametric DEA estimator formally stated as (5). The DEA output efficiency estimator of firm i is stated as $\hat{\theta}_i^{\text{DEA}} = y_i / \hat{f}^{DEA}(\mathbf{x}_i)$ and computed as

$$(\theta_i^{\text{DEA}})^{-1} = \max_{\theta\in\Re,\lambda\in\Re_+^n} \left\{ \theta \middle| \theta y_i \leq \sum\nolimits_{h=1}^n \lambda_h y_h ; \mathbf{x}_i \geq \sum\nolimits_{h=1}^n \lambda_h \mathbf{x}_h ; \sum\nolimits_{h=1}^n \lambda_h = 1 \right\} \tag{41}$$

In the second stage, the following linear equation is estimated using OLS or ML

$$\ln \hat{\theta}_i^{\text{DEA}} = \alpha + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{2-DEA},\ i = 1,...,n, \tag{42}$$

where the intercept α captures the expected inefficiency and the finite sample bias of the DEA estimator, and the composite disturbance term ε_i^{2-DEA} captures the noise term v_i and the deviations of u_i from the expected inefficiency μ. Note that the dependent variable has the “hat” because the DEA efficiency estimate is computed beforehand using (41), whereas the parameters on the right hand side of (42) are estimated using OLS or ML in a second stage.

Johnson and Kuosmanen (2012) state that the 2-DEA estimator is statistically consistent in the case of truncated noise as shown by Banker and Natarajan (2008), however, the assumptions required for consistency in Banker and Natarajan are unnecessarily restrictive.

Let $\mathbf{Z}$ denote a $n \times r$ matrix of contextual variables. Assume the noise terms are truncated as $|v_i| \leq V^M$ and denote $\mathbf{v} = (v_1,...,v_n)'$. Denote the domains of vectors $\mathbf{x}$ and $\mathbf{z}$ by D_x and D_z, respectively. Then the statistical consistency of the 2-DEA estimator can be established under the relaxed set of assumptions as follows.

Theorem 7: *If the following five assumptions are satisfied*

(i) *sequence* $\{ (y_i, \mathbf{x}_i, \mathbf{z}_i)$, i=1,...,$n\}$ *is a random sample of independent observations,*

(ii) $\lim\limits_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

(iii) *noise term* **v** *has a truncated distribution:* $|\mathbf{v}| \leq V^M \mathbf{1}$, $f_v(V^M) > 0$,

(iv) *elements of domain* D_z *are bounded from above or below such that* $\boldsymbol{\delta}'\mathbf{z}$ *has a finite maximum* $\zeta = \max\limits_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$ *at a point* $\mathbf{z}^{\xi} \in \arg\max\limits_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$,

(v) *the joint density f is continuous and satisfies* $f(\mathbf{x}, \mathbf{z}^{\xi}, 0, V^M) > 0$ *for all* $\mathbf{x} \in D_x$,

then the 2-DEA estimators are statistically consistent in the following sense

$$\operatorname*{plim}_{n\to\infty} \hat{f}^{\text{DEA}}(\mathbf{x}_i) = f(\mathbf{x}_i)\cdot \exp(V^M + \zeta) \text{ for all } i = 1,...,n,$$

$$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\delta}}^{\text{2-DEA}} = \boldsymbol{\delta}$$

Proof. See Johnson and Kuosmanen (2012), Theorem 1.

This theorem by Johnson and Kuosmanen (2012) generalizes the consistency result by Banker and Natarajan (2008) result by relaxing the following two assumptions:

1) inputs and contextual variables are statistically independent,
2) the effect of contextual variables is one-sided: $\mathbf{Z} \geq \mathbf{0}, \boldsymbol{\delta} \leq \mathbf{0}$.

Note that the DEA frontier does not converge to the true frontier f, it converges to $f(\mathbf{x})\cdot\exp(V^M+\zeta)$ (i.e., the frontier augmented by the maximum noise V^M under the ideal conditions represented by $\mathbf{z}^{\zeta}$) thus estimation of the frontier requires observing firms that are operating efficiently and are operating in the best environment and happen to get a noise drawn close to the upper bound V^M.

Consistency is a relatively weak property. In practice a data set will be finite in size and probably not as large as we would like. However, Johnson and Kuosmanen (2012) are able to provide the explicit form of the bias in the 2-DEA estimator. Specifically it depends on the bias of the DEA frontier ($\hat{f}^{\text{DEA}}$) as follows:

$$\text{Bias}(\hat{\boldsymbol{\delta}}^{\text{2-DEA}}) = -(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\left[\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X}))\right], \tag{43}$$

where

$$\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X})) = \begin{pmatrix} E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_1)) - f(\mathbf{x}_1)\cdot\exp(V^M+\zeta) \\ \vdots \\ E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_n)) - \ln f(\mathbf{x}_n)\cdot\exp(V^M+\zeta) \end{pmatrix}. \tag{44}$$

Thus, the bias of the first-stage DEA estimator carries over to the second-stage OLS regression. Importantly, the bias of the second-stage OLS estimator is due to the correlation of **Z** and bias of the first-stage DEA estimator.

In summary we would like to emphasize two critical points about 2-DEA.

1) correlation of inputs and contextual variables does not influence the statistical consistency of 2-DEA estimator as long as the columns of **X** and **Z** matrices are not linearly dependent.

2) the bias of the DEA frontier in the first-stage carries over to the second-stage OLS estimator through the correlation of the DEA frontier with the contextual variables.

We note that statistical independence of inputs and contextual variables does not necessarily guarantee that $\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X}))$ is uncorrelated with **Z**. Thus, 2-DEA does not suffer from some of the problems noted by Simar and Wilson (2011) and in fact requires significantly weaker assumptions than Banker and Natarajan (2008) suggest. However, the DEA frontier is always biased downward in a finite sample and thus this bias may be transferred to the estimation of the effect of the contextual variables. The following two sub-sections propose alternatives building on the regression interpretation of DEA which do not suffer from this bias.

7.2 One-stage DEA

The fundamental problem of the 2-DEA procedure is that the impact of the contextual variables **Z** is not taken into account in the first stage DEA. This problem has been recognized in the SFA literature, where the standard approach is to jointly estimate the frontier and the impacts of the contextual variables (e.g., Wang and Schmidt, 2002). In the similar vein, the least squares regression interpretation of DEA described in Section 4.1 allows us to estimate the DEA frontier and the coefficients $\boldsymbol{\delta}$ jointly. Specifically, we can introduce the contextual variables to the least squares formulation of DEA, stated as the QP problem (6), to obtain:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{1-DEA})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{1-DEA} \;\; \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \;\forall i \\
&\varepsilon_i^{1-DEA} \le V^M \;\; \forall i
\end{aligned} \tag{44}$$

Notable differences compared to the problem (7) concern the use of the log-transformation to enforce the multiplicative formulation of the inefficiency term (compare with Section 6.1) and the truncation of the residual ε_i^{1-DEA} at point V^M . Note that by setting $V^M = 0$ restricts the noise term to zero, and the 1-DEA formulation reduces to the joint estimation of the effect of the contextual variables and the classic deterministic DEA frontier where all input/output data is observed exactly and residuals are non-positive.

Note further that the parameter vector $\boldsymbol{\delta}$ is common to all observations, and hence it can be harmlessly omitted from the Afriat inequalities that impose convexity. In fact, the contextual variables can be interpreted as inputs that have constant marginal products across all firms[23] (i.e., we can think of matrix **Z** as a subset of **X** for which $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j \;\forall i, j$).

The statistical properties of the 1-DEA estimator generally depend on the specification of the truncation point V^M . Performance of the 1-DEA estimator has been investigated via Monte Carlo simulations in Johnson and Kuosmanen (2012) where the authors find that 1-DEA performs well even when the truncation point is misspecified. However, the assumption of truncated noise (i.e., $|v_i| \le V^M$) is non-standard and debatable (see, e.g., Simar and Wilson, 2011). While the consistency of 2-DEA critically depends on this assumption, the CNLS estimator allows us to harmlessly relax it. The next sub-section discusses the StoNED estimator with z-variables that does not rely on the truncated noise assumption.

7.3 StoNED with z-variables (StoNEZD)

Relaxing the assumption of truncated noise, we can apply CNLS to jointly estimate the expected output conditional on inputs and the effects of the contextual variables. Johnson and Kuosmanen (2011) were the first to explore this approach, referring to it as StoNED with z-variables (StoNEZD). StoNEZD incorporates the contextual variables to the stepwise procedure sescribed in Section 5. In the following, we will focus on the CNLS estimator applied in the first step: steps 2 – 4 follow as described in Section 5, and are hence omitted here.

To incorporate the contextual variables in step 1 of the StoNED estimation routine, we can refine the multiplicative CNLS problem as follows:

[23] This interpretation would vary slightly if the δ_i is negative. Then the contextual variable would be an output which would reduce the firm's ability to produce *y*.

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{CNLS} \;\; \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \;\; \forall i \\
&\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i \\
&\boldsymbol{\beta}_i \ge 0 \;\; \forall i
\end{aligned} \tag{45}$$

Note that problem (45) is identical to (44), except that the truncation constraint $\varepsilon_i \le V^M \;\forall i$ has been removed. Therefore, the least squares residuals are unrestricted, and hence problem (45) is a genuine conditional mean regression estimator.

Denote by $\hat{\boldsymbol{\delta}}^{StoNEZD}$ the coefficients of the contextual variables obtained as the optimal solution to (45). Johnson and Kuosmanen (2011) examine the statistical properties of this estimator in detail, showing its unbiasedness, consistency, and asymptotic efficiency.[24] Most importantly, the authors show that the conventional methods of statistical inference from linear regression analysis (e.g., t-tests, confidence intervals) can be applied for asymptotic inferences regarding coefficients $\boldsymbol{\delta}$. Their main result can be summarized as follows:

Theorem 8

If the following conditions are satisfied

i) *sequence* $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), i=1,...,n\}$ *is a random sample of independent observations,*

ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ *is a positive definite matrix,*

iii) *the inefficiency terms* **u** *and the noise terms* **v** *are identically and independently distributed (i.i.d.) random variables with* $Var(\mathbf{u}) = \sigma_u^2\mathbf{I}$ *and* $Var(\mathbf{v}) = \sigma_v^2\mathbf{I}$,

then the StoNEZD estimator for the coefficients of the contextual variables ($\hat{\boldsymbol{\delta}}^{StoNEZD}$) *is statistically consistent and asymptotically normally distributed according to:*

$$\hat{\boldsymbol{\delta}}^{StoNEZD} \sim_a N(\boldsymbol{\delta}, (\sigma_v^2 + \sigma_u^2)(\mathbf{Z}'\mathbf{Z})^{-1}).$$

Proof. See Johnson and Kuosmanen (2011), Theorem 2.

This theorem extends the standard result of asymptotic normality of the OLS coefficients to the StoNEZD estimator of the contextual variables. In other words, even though model (40) includes a nonparametric function in addition to a linear regression function, the presence of the nonparametric function does not affect the limiting distribution of the parameter estimator in the linear part. In addition, Johnson and Kuosmanen (2011) show that the estimator $\hat{\boldsymbol{\delta}}^{StoNEZD}$ converges at the standard parametric rate, despite the presence of the nonparametric part in the regression equation. Therefore, we can apply the standard techniques from regression analysis such as *t*-tests and confidence intervals for asymptotic inferences.

A simple trick to compute standard errors for $\hat{\boldsymbol{\delta}}^{StoNEZD}$ is to run OLS regression where the contextual variables **Z** are regressors and the dependent variable is the difference between the

[24] Johnson and Kuosmanen (2012) report some Monte Carlo evidence of the finite sample performance of the StoNEZD estimator.

natural log of observed output subtracting the natual log of the input aggregation plus 1, specifically $\ln y_i - \ln(\hat{\phi}_i + 1) = \hat{\boldsymbol{\delta}}'\mathbf{z}_i + \hat{\varepsilon}_i^{CNLS}$. This OLS regression will yield the same coefficients $\hat{\boldsymbol{\delta}}^{StoNEZD}$ that were obtained as the optimal solution to problem (45),[25] but also return the standard errors and other standard diagnostic statistics such as t-ratios, p-values, and confidence intervals.

8. Heteroscedasticity

Up to this point we have assumed that the composite error term is homoscedastic, implying the variance parameters σ_u^2 and σ_v^2 are constant across all firms. This is a standard assumption both in regression analysis and in the parametric literature of frontier estimation (e.g., Aigner et al., 1977). However, this assumption is not always realistic in applications.

We can relax the assumption of constant σ_u^2 and σ_v^2, and allow these parameters to be firm specific (i.e., $\sigma_{u,i}^2$ and $\sigma_{v,i}^2$,), and potentially dependent on inputs **x** and contextual variables **z**. We stress that the least squares approach considered in this paper enables us to apply standard econometric techniques of testing and modeling heteroscedasticity considered in the SFA literature (see, e.g., Kumbkahar et al., 1991; Caudill and Ford, 1993; Caudill et al., 1995; Battese and Coelli, 1995; Hadri, 1999; and Kumbhakar and Lovell, 2000). The purpose of this section is to provide a brief review of how some of those techniques could be adapted for the purposes of CNLS and StoNED.

The first question to consider is how would heteroscedasticity affect the CNLS and StoNED estimators if we simply ignore it? Like standard OLS, the CNLS estimator remains unbiased and consistent despite heteroscedasticity. A weighted CNLS estimator (to be considered below) might be more efficient, provided that the heteroscedastic variance parameters can be estimated with a sufficient precision. However, heteroscedasticity is not a major problem for CNLS, and trying to improve its performance through explicit modeling and estimation of heteroscedasticity may not be worth the effort. Further research would be needed to investigate this issue.

The stepwise StoNED procedure is more sensitive to heteroscedasticity, as discussed by Kuosmanen and Kortelainen (2012). At this point, we need to distinguish between i) heteroscedastic inefficiency term and ii) heteroscedasticity noise term. Ignoring type ii) heteroscedasticity is less harmful in the StoNED estimation because the skewness of the CNLS residuals is still driven by the homoscedastic inefficiency term, the expected value of inefficiency is constant, and hence the shape of the regression function (i.e., the conditional mean $E(y_i | \mathbf{x}_i)$) is identical to that of the frontier production function *f*. Type i) heteroscedasticity will cause bigger problems, as Kuosmanen and Kortelainen (2012) recognize. If the inefficiency term is heteroscedastic, then the expected value of inefficiency is no longer constant, and the shapes of the regression function and the frontier production function will diverge. To take both types of heteroscedasticity explicitly into account, in Section 8.2 we will consider a doubly-heteroscedastic model where both inefficiency and noise terms are heteroscedastic. But before proceeding to the explicit modeling of heteroscedasticity, we describe a diagnostic test of the homoscedasticity assumption.

[25] Note that this two-stage regression procedure is not subject to the problems of the 2-DEA procedure because we do control for the effects of the contextual variables in the first stage CNLS regression. It is just a computational trick to calculate the standard errors, but it can also serve as a simple diagnostic check that the solution to problem (32) is indeed optimal with respect to the contextual variables.

8.1 White test of heteroscedasticity applied to CNLS

Although the heteroscedastic inefficiency term would bias the StoNED estimator, it is important to emphasize that we do not need to take the homoscedasticity assumption by faith. Standard econometric tests of heteroscedasticity such as the White or the Breusch-Pagan tests are directly applicable to CNLS residuals. In this sub-section we briefly describe how the White (1980) test can be applied following Kuosmanen (2012).

The null hypothesis of the White test is that composite error term is homoscedastic, that is, H_0: $\sigma_{\varepsilon,i} = \sigma_{\varepsilon,j} \; \forall i, j$. The alternative hypothesis states there is heteroscedasticity, that is, H_1: $\sigma_{\varepsilon,i} \neq \sigma_{\varepsilon,j}$ for some *i,j*. Note that the alternative hypothesis does not assume any particular model of heteroscedasticity, which makes the White test compatible with the nonparametric approach. Postulating a more specific alternative hypothesis can increase the power of the test. However, the White test provides a useful starting point for more explicit modeling of heteroscedasticity.

The White test can be built upon the OLS regression of the following equation:[26]

$$(\hat{\varepsilon}_i^{CNLS})^2 = \alpha + \sum_{j=1}^{m} \beta_j x_{ij} + \frac{1}{2}\sum_{j=1}^{m}\sum_{h=1}^{j} \gamma_j x_{ij} x_{ih} + \varepsilon_i \,. \tag{46}$$

In words, we explain the squared CNLS residual by a constant, all *m* input variables, and their squared values and cross-products using a flexible quadratic functional form as an approximation of the true but unknown heteroscedasticity effects. The test statistic is

$$W = nR^2 \,,$$

where R^2 is the coefficient of determination of the OLS regression of equation (46). Under the null hypothesis of homoscedasticity, the test statistic *W* follows the $\chi^2(J)$ distribution with *J* degrees of freedom, where $J = 1 + m + m(m+1)/2$ is the number of α, β, γ parameters on the right hand side of equation (46). If the value of test statistic *W* falls below the critical value of $\chi^2(J)$ at the given level of significance (note: the usual significance levels considered are 5% and 1%), then the null hypothesis of homoscedasticity is maintained. In that case, the test result provides some additional reassurance that the original model is well specified. On the other hand, if the value of test statistic *W* exceeds the critical value of $\chi^2(J)$ at the given level of significance, then the null hypothesis is rejected, and hence explicit modeling of heteroscedasticity is needed.

The White test is usually presented in terms of the regressors of the original regression model (i.e., in terms of inputs **x** in the present context). Note that we are mainly concerned about possible heteroscedasticity with respect to inputs, which would cause bias in StoNED estimation. If we are interested in heteroscedasticity with respect to contextual variables **z**, we can also introduce the z-variables to the regression equation (46). We only need to adjust the degrees of freedom *J* to include the number of additional parameters for the z-variables, otherwise the test procedure is conducted as described above.

If significant heteroscedasticity is found, the White test does not indicate whether heteroscedasticity is in the inefficiency term or the noise term, or possibly both. To our knowledge, general diagnostic testing of whether heteroscedasticity is in the inefficiency or noise term has attracted little attention in the SFA literature. The doubly-heteroscedastic model (following Hadri, 1999; and Wang, 2002), to be examined in detail in the next sub-section, does allow us model

[26] In econometrics, heteroscedasticity is usually modeled as a function of explanatory variables (i.e., inputs **x**). In contrast, the SFA literature usually models heteroscedasticity as a function of **z**-variables that may contain some (or all) of the inputs **x**. For clarity, in this section we follow the econometric convention and focus on heteroscedasticity with respect to inputs **x** and discuss the additional **z**-variables below.

heteroscedasticity in both inefficiency and noise terms, and also test for significance of the parameter estimates. However, such specification tests are conditional on the assumed model of heteroscedasticity, including the parametric distributional assumptions regarding inefficiency and noise. An appealing feature of the White test is it does not assume any specific model of heteroscedasticity and it does not depend on the distributional assumptions. Further, the parameter estimates of the auxiliary regression (46) and the associated diagnostic tools can provide some insights on which specific inputs (or contextual variables) are most likely causes of heteroscedasticity, and whether heteroscedasticity effect appears to be linear or non-linear, and whether the interaction terms (cross-products) are significant. These insights can be useful for specifying parametric models of heteroscedasticity, to be considered in the next sub-section.

Before proceeding, note that quantile estimation (see Section 6.4) could provide a promising nonparametric route for testing heteroscedasticity. If the composite error term is homoscedastic, then the *q*-quantiles should have approximately same shape for different values of parameter *q*. Provided that the number of input (and output) variables is sufficiently small, plotting the estimated *q*-quantiles at different values of *q* allow one to visually inspect whether homoscedasticity holds by a reasonable approximation. If homoscedasticity is violated, the *q*-quantile plots can help one to identify in which part of the frontier heteroscedasticity occurs, and which inputs are likely sources of heteroscedasticity. In the context of linear quantile regression, Koenker and Bassett (1982) propose formal tests of heteroscedasticity based on the comparison of the estimated *q*-quantiles at different values of *q*. Newey and Powell (1987) apply a similar idea for the *q*-expectiles, noting that the *q*-expectiles could also be used for testing symmetry of the composite error term (i.e., whether the asymmetric inefficiency term *u* is significant; compare with Section 5.5). Adapting these tests to the nonparametric CNQR method for estimating *q*-quantiles and the CAWLS method for estimating *q*-expectiles introduced in Section 6.4 provides an interesting challenge for future research further discussed in section 9.

8.2 Doubly-heteroscedastic model

If the White test indicates significant heteroscedasticity, it is difficult to tell *a priori* whether heteroscedasticity is due to the inefficiency term, the noise term, or possibly both. Therefore, we will consider the general doubly-heteroscedastic model where both the inefficiency and noise term can be heteroscedastic. The doubly-heteroscedastic model was first considered by Hadri (1999). Our formulation below is mainly based on Wang (2002) and Kumbhakar and Sun (2013).

Consider the unified model described in Section 2. In this section we assume the inefficiency term has a truncated normal distribution and the noise term is normally distributed according to

$$u_i \sim N^+(\mu_i, \sigma_{u,i}^2)$$

$$v_i \sim N(0, \sigma_{v,i}^2)$$

The pre-truncation mean of the inefficiency term is assumed to be a linear function of inputs:

$$\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i .$$

The pre-truncation standard deviation of the inefficiency term and the standard deviation of the noise term are specified as

$$\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$$

$$\sigma_{v,i} = \exp(\alpha_2 + \boldsymbol{\rho}'\mathbf{x}_i)$$

Note that the exponent functions are commonly used in this context to guarantee that the standard deviations are positive at all input levels. While the specific parametric assumption may appear arbitrary, this model is one of the most flexible and general parametric specifications of heteroscedasticity. Note that the truncated normal distribution where both the pre-truncation mean and variance depend on the input level allows that the location (mean) and the shape (variance)of the inefficiency distribution can change as a function of inputs.

This formulation of heteroscedastic inefficiency term implies that the expected value of inefficiency can be stated as (see Wang, 2002; Kumbhakar and Sun, 2013)

$$E(u_i \mid u_i > 0) = \sigma_{u,i} \left[\Lambda_i + \frac{\phi(\Lambda_i)}{\Phi(\Lambda_i)} \right], \tag{47}$$

where

$$\Lambda_i = \frac{\mu_i}{\sigma_{u,i}}$$

and ϕ and Φ are the density function and the cumulative distribution function of the standard normal $N(0,1)$ distribution, respectively. The expected inefficiency is no longer a constant, but its dependence on inputs **x** has a well-defined functional form conditional on the parametric assumptions stated above. This allows us to both estimate the heteroscedasticity effects empirically, and take heteroscedasticity explicitly into account in the StoNED procedure.

8.3 Stepwise StoNED estimation under heteroscedasticity

To estimate the doubly-heteroskedastic model, we can adjust the stepwise StoNED routine presented in Section 5 as follows (a more detailed elaboration of each step follows below):

Step 1: Apply the CNLS estimator (3) to estimate the conditional mean output $\hat{g}^{CNLS}(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i)$.

Step 2: Apply quasi-likelihood estimation to the CNLS residuals ε_i^{CNLS} to estimate the parameters of μ_i, $\sigma_{u,i}$, and $\sigma_{v,i}$.

Step 3: Adjust the conditional mean function by adding the expected inefficiency $E(u_i \mid \mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i})$ to estimate the frontier for the observed data points using

$$\hat{f}^{StoNED}(\mathbf{x}_i) = \hat{g}^{CNLS}(\mathbf{x}_i) + E(u_i \mid \mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i}).$$

Then apply equation (5) to estimate the frontier $\hat{f}_{\min}^{StoNED}(\mathbf{x})$ for unobserved points.

Step 4: Apply JLMS method to estimate firm-specific inefficiency using the conditional mean $E(u_i \mid \hat{\varepsilon}_i^{CNLS})$.

In step 1, we estimate the conditional mean function $g(\mathbf{x})$. The CNLS estimator remains unbiased and consistent estimator of the conditional mean g, despite heteroscedastic composite error term (similar to OLS). However, note that in the case of the doubly-heteroscedastic model

$$g(\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i) = f(\mathbf{x}_i) - E(u_i \mid \mathbf{x}_i).$$

Note that the shape of function g can differ from that of frontier f because $E(u_i \mid \mathbf{x}_i)$ is a function of inputs **x**. We will take this into account in step 3 where we shift function g upward, not by a constant μ, but rather, by the estimated $E(u_i|\mathbf{x}_i)$.[27] It is also worth noting that function g is not necessarily monotonic increasing and concave even if the production function f satisfies these axioms because $-E(u_i|\mathbf{x}_i)$ can be a non-monotonic and non-concave function of inputs (note: there does exist parameter values for which $-E(u_i|\mathbf{x}_i)$ is indeed monotonic and concave in the domain of non-negative **x**). To apply CNLS in step 1, we need to assume that the curvature of the production function f dominates and that function g is monotonic increasing and concave (at least by approximation). Even if one assumes that f exhibits CRS, it is recommended to apply the VRS specification in step 1 to allow for the nonlinear effects of $E(u_i|\mathbf{x}_i)$, and impose CRS later in step 3.

Having estimated the parameters of the inefficiency and noise terms, it is possible to test if monotonicity and concavity assumptions of g hold. If g does not satisfy monotonicity and concavity, we can substitute CNLS by techniques depending on which axiom does not hold. Specifically, if the concavity assumption is violated, it is possible to apply isotonic nonparametric least squares (INLS) suggested by Keshvari and Kuosmanen (2013). Another possibility is to estimate order-q quantile frontier using either CNQR or CAWLS techniques introduced in Section 6.4. Specifying the correct value for q will ensure that the quantile frontier inherits the monotonicity and concavity properties of frontier f even if the heteroscedastic inefficiency term is a non-monotonic or non-convex function of inputs. Indeed, we do not insist on estimating the conditional mean in step 1, the conditional quantile is equally suitable.

In step 2 it is natural to resort to the pseudolikelihood method since we utilize a rather heavily parametrized model of heteroscedasticity. As already noted in Section 5, a simple practical trick to conduct quasi-likelihood estimation is to use the standard ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). In this case we specify the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), and the ML algorithm performs the quasilikelihood estimation. For example, the frontier modeling tools of Stata allows one to include "explanatory variables for technical inefficiency variance function (uhet)" and "explanatory variables for idiosyncratic error variance function (vhet)" if the distribution of inefficiency term is specified as half-normal or exponential. It is also possible to include covariates to the truncated normal specification of the inefficiency term, but in this specification the noise term is assumed to be homoscedastic. Hung-Jen Wang has developed a Stata package for the model described in Wang (2002), which can be used for estimating the model estimating the heteroscedasticity model described above.[28]

Having estimated the underlying parameters of $\mu_i, \sigma_{u,i}, \sigma_{v,i}$, it is recommended to apply standard specification tests available for ML (i.e., likelihood-ratio, Lagrange multiplier, or Wald test) to test restrictions $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\gamma} = \mathbf{0}$, and $\boldsymbol{\rho} = \mathbf{0}$. For example, if the null hypothesis of $\boldsymbol{\rho} = \mathbf{0}$ is not rejected, then the assumption of homoscedastic noise term can be maintained. Similarly, if $\alpha_0 = 0$, $\boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\gamma} = \mathbf{0}$, then the model of heteroscedastic truncated normal inefficiency term reduces to a homoscedastic half-normal inefficiency term. If the specification tests provide evidence that some

[27] In the context of SFA, Kumbhakar and Lovell (2000) state strongly that the stepwise MOLS procedure cannot be used in the case of heteroscedastic inefficiency. They correctly note that the OLS estimator used in the first step yields biased estimates of not only the intercept but also the slope coefficients of the frontier. However, Kumbhakar and Lovell seem to overlook the possibility of eliminating the bias by shifting function g upward by a conditional expectation of inefficiency that depends on inputs **x**.

[28] The Stata package is available from Wang's homepage: http://homepage.ntu.edu.tw/~wangh/.

of the heteroscedasticity effects are not significant, we would recommend excluding those effects from the heteroscedasticity model and estimating step 2 again.

One additional issue is in the context of linear regressionthat efficiency of the least squares estimator can be improved by applying weighted least squares or generalized least squares. Having estimated the firm specific $\sigma_{u,i}, \sigma_{v,i}$, it is possible to return back to step 1 and apply a weighted version of the CNLS estimator. Defining $\hat{\sigma}^2_{\varepsilon,i} = \hat{\sigma}^2_{u,i} + \hat{\sigma}^2_{v,i}$, we can modify the objective function of the CNLS problem as

$$\min \sum_{i=1}^{n} \frac{(\varepsilon_i^{CNLS})^2}{\hat{\sigma}^2_{\varepsilon,i}}$$

maintaining the original constraints of (3). Interpreting the given $1/\hat{\sigma}^2_{\varepsilon,i}$ as firm-specific weights, this weighted least squares formulation of CNLS is directly analogous to the generalized least squares (GLS) estimator of the linear regression model.[29] However, as yet there is no evidence that the use of weighted least squares can improve efficiency of the CNLS estimator. Intuitively, the direct analogue with GLS would suggest that weighted least squares can be more efficient than the unweighted CNLS under heteroscedasticity. On the other hand, recall that CNLS approximates the underlying function g by a piece-wise linear curve. Since the hyperplane segments of the unweighted CNLS formulation provide local approximation, assigning larger or smaller weights to certain regions of the frontier may not have much effect on the piece-wise linear approximation. In our limited experience, introducing the weights $1/\hat{\sigma}^2_{\varepsilon,i}$ does not necessarily have any notable impact on the results. Further, we need to be able to estimate $\sigma^2_{\varepsilon,i}$ with a sufficient precision. Overall, we are somewhat skeptical whether the possible benefit in terms of improved efficiency of the CNLS estimator can outweigh the cost of additional effort of conducting the weighted least squares estimation. This forms an interesting open question for future research.

In step 3 we adjust the conditional mean function g estimated in step 1 (or alternatively, the conditional q-quantile) for the estimated expected inefficiency to estimate the frontier f. Note that the conditional mean $E(u_i|\mathbf{x}_i)$ is no longer a constant, but a function that depends on inputs **x**. Using equation (47), we can write the estimated expected inefficiency as the function of inputs and parameter estimates as

$$\begin{aligned} E(u_i|\mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i}) &= \hat{\mu}_i + \hat{\sigma}_{u,i}\frac{\phi(\hat{\Lambda}_i)}{\Phi(\hat{\Lambda}_i)} \\ &= (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i) + \exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)\left[\phi\left(\frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)}\right)\Big/\Phi\left(\frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)}\right)\right] \end{aligned}$$

This expression reveals that in the doubly-heteroscedastic model the expected value of inefficiency has a linear part originating from the mean $\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i$, and a nonlinear part driven by $\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$. Having estimated the parameters of the inefficiency term, it is useful to evaluate whether $-E(\hat{u}_i|\mathbf{x}_i)$ is monotonically increasing and concave within the observed range of inputs (e.g., plot the values of $-E(\hat{u}|\mathbf{x})$ at different levels of **x** to visually inspect possible violations of monotonicity and concavity). To ensure that the estimated frontier function satisfies the

[29] Note that in the CNLS context we prefer to introduce weights to the objective function instead of applying variable transformations (as in GLS) because the monotonicity and concavity constraints must hold for the original input variables **x**.

postulated axioms despite minor violations of monotonicity and concavity (which may be just artifacts of the arbitrary parametric specification of the heteroscedasticity model), we apply the minimum extrapolation principle and utilize the DEA method stated in equation (5) to obtain the convex monotonic hull of the fitted values $\hat{f}^{StoNED}(\mathbf{x}_i)$ of observations i=1,…,n, which yields the frontier estimator $\hat{f}_{\min}^{StoNED}(\mathbf{x})$.

In step 4, we can compute firm specific inefficiency estimates using the JLMS conditional mean $E(u_i \,|\hat{\varepsilon}_i^{CNLS})$ using the firm specific parameter estimates $\hat{\mu}_i, \hat{\sigma}_{u,i}, \hat{\sigma}_{v,i}$. Note that the expected inefficiency $E(u_i \,|\mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i})$ applied for shifting the conditional mean function g to estimate frontier f does not depend on the heteroscedasticity of the noise term. However, the JLMS efficiency does also depend on the heteroscedasticity of the noise term $\hat{\sigma}_{v,i}$. Kumbhakar and Sun (2013) discuss this issue in more detail, showing that the marginal effect of inputs on the conditional JLMS efficiency also depend on the heteroscedasticity of the noise term.

9. Directions for future research

This chapter has provided an updated and elaborated presentation of the CNLS and StoNED methods. Bridging the gap between the established DEA and SFA paradigms, these methods represent a major paradigm shift towards a unified and integrated methodology of frontier estimation and efficiency analysis that has a considerably broader scope than the conventional DEA and SFA tools. This chapter did not only review previously published method developments and their extensions, but also presented some new innovations, including the first extension of the StoNED method to the general case of multiple inputs and multiple outputs, and the first detailed examination of how heteroscedastic inefficiency and noise terms can be modeled within the CNLS and StoNED estimation frameworks.

We see CNLS and StoNED not only as the state of the art in axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise, but also a promising way forward. Kuosmanen and Kortelainen (2012) stated explicitly 12 promising avenues of future research on the StoNED methodology. In the following we will provide an updated version of a 12 point research program, indicating the work that has already been done as well as work that remains to be done.

”1. *Adapting the known econometric and statistical methods for dealing with heteroskedasticity, endogeneity, sample selection, and other potential sources of bias, to the context of CNLS and StoNED estimators*.”

In this chapter we presented the first detailed examination about the modeling of heteroscedasticity in the inefficiency and noise terms. Kuosmanen, Johnson and Parmeter (2013) examine the endogeneity problem from a novel perspective employing directional distance functions. Obviously, a lot of further work is needed in this area. Alternative models of heteroscedasticity as well as estimation techniques deserve careful attention. The convex nonparametric quantile regression and the convex asymmetrically weighted least squares methods discussed in Section 6.4 and the generalized least squares estimator discussed in Section 8.3 provide potential methods for modeling and testing heteroskedasticity. The use of instrumental variables in CNLS for modeling measurement errors, sample selection, and other types of endogeneity bias should be investigated.

“2. *Extending the proposed approach to a multiple output setting.*”

In this chapter we also presented the first extension of the StoNED method to the general case of multiple inputs and multiple outputs using the directional distance function (see also Kuosmanen, Johnson and Parmeter, 2013). Further work is also needed in this area. Alternative representations of the joint production technology, including the radial input and output distance functions, should be investigated. The main challenge in modeling joint production is not the formulation of the mathematical programming problem for the CNLS estimator (the usual DEA problem) or deconvoluting the composite error term (the usual SFA problem). The main challenge is the probabilistic modeling of the data generating process in the case of joint production, involving multiple endogenous inputs and outputs. Kuosmanen, Johnson and Parmeter (2013) provides a useful starting point in this respect.

“3. *Extending the proposed approach to account for relaxed concavity assumptions (e.g., quasiconcavity)*.”
Keshvari and Kuosmanen (2013) presented the first extension in this direction, applying isotonic regression that relaxes the concavity assumption of CNLS. This approach estimates a step function analogous to free disposable hull (FDH) in the middle of the data cloud. The insights of Keshvari and Kuosmanen could be useful for examining the intermediate cases between the non-convex step function and the fully convex CNLS, allowing one to postulate quasiconcavity or quasiconvexity in terms of some variables (e.g., inputs, or input prices in the estimation of the cost function). Many opportunities for future research exist in this direction.

“4. *Developing more efficient computational algorithms or heuristics for solving the CNLS problem*.”
Lee et al. (2013) is the first contribution in this direction. The algorithm developed in that paper first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. We believe the computational efficiency can be improved considerably by clever algorithms and heuristics (see, e.g., Hannah and Dunson, 2013). This is an important avenue for future research in the era of “big data”.

“5. *Examining the statistical properties of the CNLS estimator, especially in the multivariate case*.”
Seijo and Sen (2011) and Lim and Glynn (2012) were the first to address this challenge, proving statistical consistency of the CNLS estimator in the general multivariate case under slightly different assumptions about the data generating process. Further research on both the finite sample properties (e.g., unbiasedness or bias, efficiency, mean squared error) and the asymptotic properties (e.g., rates of convergence, limiting distributions) under different assumption of the data generating process would be needed. In this respect, Groeneboom et al. (2001a,b) provide an excellent starting point. The statistical properties of the convex nonparametric quantile regression (CNQR) and the convex asymmetrically weighted least squares (CAWLS) methods introduced in Section 6.4 also deserve further research.

“6. *Investigating the axiomatic foundation of the CNLS and StoNED estimators*.”
CNLS regression builds upon the same axioms as DEA, and StoNED estimation applies the minimum extrapolation principle to obtain a unique frontier function that satisfies the postulated axioms. However, it would be compelling if the technology characterized by CNLS and/or StoNED could be stated rigorously from the axiomatic point of view as the intersection of all sets that satisfy

the stated axioms and satisfy axiom X. It remains unknown whether axiom X exists, and how it could be formulated explicitly.

"7. *Implementing alternative distributional assumptions and estimating the distribution of the inefficiency term by semi- or nonparametric methods in the cross-sectional setting*."

In this chapter (Section 5.2) we have provided an extensive review of possibilities, including parametric and semi-parametric alternatives. In principle, the quasilikelihood method is applicable to any parametric specification of inefficiency distribution. The most promising way forward seems to be the nonparametric kernel deconvolution of the CNLS residuals, following the works by Hall and Simar (2002) and Horrace and Parmeter (2011). One challenge that remains is to adapt the JLMS conditional mean inefficiency to the semi-parametric setting where no parametric distribution is specified for the inefficiency term.

"8. *Distinguishing time-invariant inefficiency from heterogeneity across firms, and identifying inter-temporal frontier shifts and catching up in panel data models*."

Kuosmanen and Kortelainen (2012) present a simple fixed effects approach to modeling panel data, assuming time-invariant inefficiency. In this chapter we considered the parallel random effects approach, following Eskelinen and Kuosmanen (2013). Ample opportunities for extending these basic techniques to more sophisticated semi-parametric models allowing for technical progress and time-varying inefficiency are available. Indeed, panel data models have been extensively studied both in general econometrics and in the SFA literature. Both the insights and practical solutions from panel data econometrics can be imported to the CNLS and StoNED framework.

"9. *Extending the proposed approach to the estimation of cost, revenue, and profit functions as well as to distance functions*."

Kuosmanen and Kortelainen (2012) consider the estimation of cost function in the single output case under CRS. They made these restrictive assumptions because the cost function must be a concave function of input prices. However, if the standard convexity axiom of the production possibility set holds, then the cost function is a convex function of outputs. A challenge that remains is to formulate the CNLS problem such that we can estimate a function that is convex in one subset of variables (i.e., outputs), but concave in another subset of variables (i.e., input prices). Kuosmanen (2012) estimates a multi-output cost function using StoNED, but the input prices were excluded by assuming that all firms take the same input prices as given.

"10. *Developing a consistent bootstrap algorithm and/or other statistical inference methods*."

An earlier version of Kuosmanen and Kortelainen (2012) proposed to adapt the parametric bootstrap method proposed by Simar and Wilson (2010) for drawing statistical inferences in the StoNED setting. However, the anonymous reviewers were not convinced that the proposed boostrap method is necessarily consistent when applied to the CNLS residuals. Indeed, one should be wary of naïve bootstrap and resampling approaches that produce invalid and misleading results. Since Kuosmanen and Kortelainen were not able to prove consistency of Simar and Wilson's bootstrap procedure in the CNLS case, the suggestion was excluded from the published version. We stress that adapting one of the known variants of the bootstrap method to the context of CNLS and StoNED would be straightforward. The challenge is to prove that the chosen version of bootstrap method is consistent under the stated assumptions about the data generating process. Another promising approach is to test if CNLS estimates differ significantly from the corresponding estimates obtained using parametric methods (see Sen and Meyer, 2013). As for the contextual variables, Johnson and Kuosmanen (2012) prove that conventional inference techniques from linear regression analysis (e.g., t-tests, p-values, confidence intervals) can be applied for the parametric part (i.e., the coefficients of the contextual variables).

"11. *Conducting further Monte Carlo simulations to examine the performance of the proposed estimators under a wider range of conditions, and comparing the performance with other semi- and nonparametric frontier estimators*."
Several published studies provide Monte Carlo evidence on the finite sample performance of CNLS and StoNED estimators. Kuosmanen (2008) and Kuosmanen and Kortelainen (2012) provide the first simulation results for CNLS and StoNED, respectively, focusing on the precision in estimating the frontier production function f. Johnson and Kuosmanen (2011) present MC simulations regarding the estimation of the parametric δ representing the effect of a single contextual variable z that may be correlated with input x. Andor and Hesse (in press) provide an extensive comparison of the performances of DEA, SFA, and StoNED, mainly focusing on the estimation of the firm specific inefficiency u_i. However, note that all estimators considered are inconsistent in the noisy setting considered because u_i is just a single realization of a random variable. Kuosmanen, Saastamoinen and Sipiläinen (2013) compare performances of DEA, SFA and StoNED in terms of estimating a frontier cost function. They calibrate their simulations to match the empirical characteristics of the Finnish electricity distribution firms. Their simulations demonstrate that if the premises stated by the Finnish energy regulator hold, then the StoNED estimator has superior performance compared to its restricted special cases, DEA and SFA. As for further research, it would be interesting to compare performance of CNLS and StoNED with those of other semi- and nonparametric frontier estimation techniques such as kernel regression and local maximum likelihood.

"12. *Applying the proposed method to empirical data, and adapting the method to better serve the needs of specific empirical applications*."
The first published application of the StoNED method was Kuosmanen and Kuosmanen (2009), who estimated the production function from the data of 332 Finnish dairy farms in order to assess sustainability performance of farms. Subsequently, there have been several applications in the energy sector, both in production and distribution of electricity. Mekaroonreung and Johnson (2012) applied StoNED to estimate the shadow prices of SO2 and NOx from the data of U.S. coal-fired power plants. Thus far, the most significant real-world application of StoNED has been the study by Kuosmanen (2012) [see also Kuosmanen, Saastamoinen and Sipiläinen (2013), Dai and Kuosmanen (2014), and Saastamoinen and Kuosmanen (2014)]. Based on the results of this study, the Finnish energy market regulator adopted the StoNED method in systematic use in the regulation of the Finnish electricity distribution industry, with the total annual turnover of more than €2 Billion. Another real-world application of StoNED is Eskelinen and Kuosmanen (2013), who assessed inter-temporal performance of sales teams using monthly data of Helsinki OP-Pohjola Bank, in close collaboration with the central management of the bank. The results and insights gained in this study were communicated to the team managers and were utilized for setting performance targets for sales teams. These empirical applications illustrate the flexibility and adaptability of the StoNED methodology to suit the specific needs of the application. The applications also provide motivation for developing further methodological extensions to meet the requirements of future applications.

In conclusion, we hope the 12-point program discussed above might inspire future methodological research along the lines described or along new avenues that have escaped our attention. We also hope that the methodological tools currently available would find inroads to empirical applications. In our experience from both Monte Carlo simulations and real empirical applications, CNLS and StoNED has proved dependable, reliable and robust, with an ability to produce results and insights that could not be found using the conventional methods.

References

Afriat, S.N. (1967). The Construction of a Utility Function from Expenditure Data. *International Economic Review* 8: 67-77.

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

Aigner, D. and S. Chu (1968). On estimating the industry production function. *American Economic Review* 58: 826-839.

Aigner, D., Lovell, C.A.K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21-37.

Alminidis, P., J. Qian and R. Sickles (2009). Stochastic Frontiers with Bounded Inefficiency, *mimeo*, Rice University.

Almanidis, P. and R. C. Sickles (2011). The skewness issue in stochastic frontier models: Fact of fiction? In I. van Keilegom and P. W. Wilson (Eds.), *Exploring Research Frontiers in Contemporary Statistics and Econometrics*. Springer Verlag, Berlin Heidelberg.

Andor, M. and F. Hesse (2014). The StoNED Age: The Departure Into a New Era of Efficiency Analysis? – A Monte Carlo Comparison of StoNED and the "Oldies" (SFA and DEA). *Journal of Productivity Analysis* 41(1), 85-109.

Aragon, Y., A. Daouia, and C. Thomas-Agnan (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory* 21: 358–389.

Banker, R.D. (1993). Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation. *Management Science* 39, 1265-1273.

Banker, R.D., A. Charnes, W.W. Cooper. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science* 30(9): 1078-1092.

Banker, R.D., and A. Maindiratta (1986). Piece-Wise Loglinear Estimation of Efficient Production Surfaces. *Management Science* 32(1): 126-135.

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

Banker, R.D. and R. Natarajan (2008). Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research* 56(1): 48-58.

Battese, G.E. and T.J. Coelli, (1995). A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data, *Empirical Economics* 20(2): 325-332.

Carree, M.A. (2002). Technological inefficiency and the skewness of the error component in stochastic frontier analysis. *Economics Letters* 77: 101–107.

Caudill, S., and J. Ford (1993). Biases in frontier estimation due to heteroscedasticity. *Economics Letters* 41(1): 17-20.

Caudill, S., J. Ford , and D. Gropper (1995). Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. *Journal of Business and Economic Statistics* 13(1): 105-111.

Chambers, R.G., Y.H. Chung, and R. Färe (1996). Benefit and distance functions. *Journal of Economic Theory* 70(2): 407-419.

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

Charnes, A., W.W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429-444.

Charnes, A., W.W. Cooper, L. Seiford, and J. Stutz (1982). A multiplicative model for efficiency analysis. *Socio-Economic Planning Sciences* 16: 223–224.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Vol. 6*. North Holland.

Cobb, C.W. and P.H. Douglas (1928). A Theory of Production. *American Economic Review* 18: 139-165.

Coelli, T. (1995), Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. *Journal of Productivity Analysis* 6, 247-268.

Coelli, T., and S. Perelman (1999). A comparison of parametric and non-parametric distance functions: With application to European railways. *European Journal of Operational Research* 117(2): 326-339.

Coelli, T., and S. Perelman (2000). Technical efficiency of European railways: a distance function approach. *Applied Economics* 32(15): 1967-1976.

D'Agostino, R., and E.S. Pearson (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b2 and $\sqrt{b_1}$. *Biometrika* 60(3): 613-622.

Dai, X. and T. Kuosmanen (2014). Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega* 42(1): 179-188.

Dantzig, G.B., D.R. Fulkerson, and S.M. Johnson (1954). Solution of a large-scale traveling salesman problem. *Operations Research* 2, 393–410.

Dantzig, G.B., D.R. Fulkerson, and S.M. Johnson (1959). On a linear-programming combinatorial approach to the traveling-salesman problem. *Operations Research* 7, 58–66.

Daouia, A., and L. Simar (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics* 140: 375–400.

Eskelinen, J. and T. Kuosmanen (2013). Intertemporal efficiency analysis of sales teams of a bank: Stochastic semi-nonparametric approach. *Journal of Banking and Finance* 37(12): 5163-5175.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272

Fan, Y., Q. Li, and A. Weersink (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business and Economic Statistics* 14, 460-468.

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society Series A* 120: 253-281.

Färe, R., S. Grosskopf, D.-W. Noh, and W. Weber (2005). Characteristics of a polluting technology: theory and practice. *Journal of Econometrics* 126: 469-492.

Färe, R., S. Grosskopf, M. Norris and Z. Zhang (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review* 84(1): 66–83.

Gabrielsen, A. (1975). On estimating efficient production functions. Working Paper No. A-85, Chr. Michelsen Institute, Department of Humanities and Social Sciences, Bergen, Norway.

Greene, W.H. (1980). Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* 13: 26-57.

Greene, W.H. (2008) The econometric approach to efficiency analysis. In H.O. Fried, C.A.K. Lovell, and S.S. Schmidt (Eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 92-250). New York, Oxford University Press Inc.

Groeneboom, P., G. Jongbloed, and J.A. Wellner (2001a). A canonical process for estimation of convex functions: the ‘‘Invelope’’ of integrated brownian motion +t4. *Annals of Statistics* 29:1620–1652.

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

Gstach, D. (1998). Another approach to data envelopment analysis in noisy environments: DEA+. *Journal of Productivity Analysis* 9(2): 161-176.

Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics* 17 (3), 359-363.

Hall, P., and L. Simar (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association* 97: 523-534.

Hannah, L.A. and D.B. Dunson (2013). Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research* 14: 3207–3240.

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

Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44(3): 461-465.

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

Hoff, A. (2007). Second stage DEA: Comparison of approaches for modeling the DEA score. *European Journal of Operational Research*, 181, 425-435.

Horrace, W., and C. Parmeter (2011). Semiparametric deconvolution with unknown error variance. *Journal of Productivity Analysis* 35(2): 129-141.

Johnson, A.L., and T. Kuosmanen (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. *Journal of Productivity Analysis* 36 (2), 219-230.

Johnson, A.L., and T. Kuosmanen (2012). One-stage and two-stage DEA estimation of the effects of contextual variables. *European Journal of Operational Research* 220, 559-570.

Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt (1982). On estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233-238.

Keshvari, A. and T. Kuosmanen (2013). Stochastic non-convex envelopment of data: Applying isotonic regression to frontier estimation. *European Journal of Operational Research* 231, 481-491.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R., and G.W. Bassett (1978). Regression quantiles. *Econometrica* 46(1): 33–50.

Koenker, R., and G.W. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles, *Econometrica* 50: 43-61.

Krugman, P. (1992). *The age of diminished expectations: US economic policy in the 1980s*, MIT Press, Cambridge.

Kumbhakar, S.C., S. Ghosh, and J.T. McGuckin (1991). A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *Journal of Business and Economic Statistics* 9(3): 279-286.

Kumbhakar, S.C., and C.A.K. Lovell (2000). *Stochastic frontier analysis*. New York, USA, Cambridge University Press.

Kuosmanen, T. (2006): Stochastic nonparametric envelopment of data: Combining virtues of SFA and DEA in a unified framework, MTT Discussion Paper No. 3/2006.

Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. *Econometrics Journal* 11, 308-325.

Kuosmanen, T. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics* 34: 2189-2199.

Kuosmanen, T., and M. Fosgerau (2009). Neoclassical versus frontier production models? Testing for the skewness of regression residuals. *Scandinavian Journal of Economics* 111(2): 351-367.

Kuosmanen, T., and N. Kuosmanen (2009). Role of benchmark technology in sustainable value analysis: An application to Finnish dairy farms. *Agricultural and Food Science* 18 (3-4), 302-316.

Kuosmanen, T., and A.L. Johnson (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research,* 58, 149-160.

Kuosmanen, T., A.L. Johnson, and C. Parmeter (2013). Orthogonality conditions for identification of joint production technologies: Axiomatic nonparametric approach to the estimation of stochastic distance functions, unpublished working paper (available from the authors by request).

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

Kuosmanen, T., A. Saastamoinen, and T. Sipiläinen (2013). What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. *Energy Policy* 61: 740-750.

Lee, C-Y, A.L. Johnson, E. Moreno-Centeno, and T. Kuosmanen (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research* 227(2):391-400.

Lim, E., and P.W. Glynn (2012). Consistency of multidimensional convex regression. *Operations Research* 60(1), 196-208.

Lovell, C.A.K., S. Richardson, P. Travers, and L.L. Wood (1994). Resources and functionings: A new view of inequality in Australia (with), in W. Eichhorn, ed., *Models and measurement of welfare and inequality*, pp. 787-807, Springer, Berlin, Heidelberg, New York.

Marschak, J., and W. Andrews (1944). Random simultaneous equations and the theory of production, *Econometrica* 12, 143–205.

McDonald, J. (2009). Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research* 197, 792-798.

Meeusen, W., and J. Vandenbroeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18(2): 435-445.

Mekaroonreung, M., and A.L. Johnson (2012). Estimating the shadow prices of SO2 and NOx for U.S. coal power plants: A convex nonparametric least squares approach. *Energy Economics* 34(3): 723-732.

Newey, W.K., and J.L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica* 55(4): 819-847.

Ondrich, J., and J. Ruggiero (2001). Efficiency measurement in the stochastic frontier model. *European Journal of Operational Research* 129, 434-442.

Ruggiero, J. (2004). Data envelopment analysis with stochastic data. *Journal of the Operational Research Society* 55(9): 1008-1012.

Saastamoinen, A., and T. Kuosmanen (2014). Quality Frontier of Electricity Distribution: Supply Security, Best Practices, and Underground Cabling in Finland, *Energy Economics*, in press. DOI: 10.1016/j.eneco.2014.04.016.

Schmidt, P., and T. Lin, (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometric*, 24: 349-361.

Schmidt, P., and R.C. Sickles (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics* 2(4): 367-74.

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

Sen, B., and M. Meyer (2013). Testing against a parametric regression function using ideas from shape restricted estimation, arXiv preprint arXiv:1311.6849, available at: http://arxiv.org/pdf/1311.6849.pdf.

Simar, L., and P.W. Wilson (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44(1): 49-61.

Simar, L., and P.W. Wilson (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* 27(6): 779-802.

Simar, L., and P.W. Wilson (2007). Estimation and inference in two-stage, semi-parametric models of production processes, *Journal of Econometrics* 136(1): 31-64.

Simar, L., and P.W. Wilson (2010). Inferences from cross-sectional, stochastic frontier models. *Econometric Reviews* 29(1): 62-98.

Simar, L., and P.W. Wilson (2011). Two-stage DEA: Caveat emptor. *Journal of Productivity Analysis* 36(2): 205-218.

Timmer, C.P. (1971). Using a probabilistic frontier production function to measure technical efficiency. *Journal of Political Economy* 79: 767-794.

Varian, H.R. (1984). The nonparametric approach to production analysis. *Econometrica* 52, 579-598.

Verbeek, M. (2008). *A Guide to Modern Econometrics*. England, John Wiley & Sons Ltd.

Wang, H., and P. Schmidt (2002). One step and two step estimation of the effects of exogenous variables on technical efficiency levels. *Journal of Productivity Analysis* 18, 129-144.

Wang, Y., S. Wang, C. Dang, and W. Ge (2014). Nonparametric quantile frontier estimation under shape restriction. *European Journal of Operational Research* 232: 671-678.

Winsten, C.B. (1957). Discussion on Mr. Farrell's Paper. *Journal of the Royal Statistical Society Series A* 120(3): 282-284.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4): 817–838.