Estimation of semiparametric stochastic frontiers under shape constraints with application to pollution generating technologies

Kortelainen, Mika

20 June 2008

Estimation of Semiparametric Stochastic Frontiers Under Shape Constraints with Application to Pollution Generating Technologies

Mika Kortelainen*
FDPE and University of Joensuu

Abstract

A number of studies have explored the semi- and nonparametric estimation of stochastic frontier models by using kernel regression or other nonparametric smoothing techniques. In contrast to popular deterministic nonparametric estimators, these approaches do not allow one to impose any shape constraints (or regularity conditions) on the frontier function. On the other hand, as many of the previous techniques are based on the nonparametric estimation of the frontier function, the convergence rate of frontier estimators can be sensitive to the number of inputs, which is generally known as "the curse of dimensionality" problem. This paper proposes a new semiparametric approach for stochastic frontier estimation that avoids the curse of dimensionality and allows one to impose shape constraints on the frontier function. Our approach is based on the single-index model and applies both single-index estimation techniques and shape-constrained nonparametric least squares. In addition to production frontier and technical efficiency estimation, we show how the technique can be used to estimate pollution generating technologies. The new approach is illustrated by an empirical application to the environmental adjusted performance evaluation of U.S. coal-fired electric power plants.

JEL Classification: C14, C51, D24, Q52

Key Words: stochastic frontier analysis (SFA), nonparametric least squares, single-index model, sliced inverse regression, monotone rank correlation estimator, environmental efficiency

---

* Finnish Doctoral Programme in Economics (FDPE) and Department of Economics and Business Administration, University of Joensuu, P.O. Box 111, 80101 Joensuu, Finland. Email: mika.kortelainen@joensuu.fi

1. Introduction

Estimation of production frontiers is usually based either on the nonparametric data envelopment analysis (DEA: Farrell, 1957; Charnes et al. 1978) or on the parametric stochastic frontier analysis (SFA: Aigner et al., 1977; Meeusen and van den Broeck, 1977). While traditional SFA builds on parametric regression techniques, DEA is based on a linear programming formulation that does not assume a parametrical functional form for the frontier, but relies on general regularity properties such as monotonicity and convexity. Although both DEA and SFA have their own weaknesses, it is generally accepted that the main appeal of SFA is its stochastic, probabilistic treatment of inefficiency and noise, whereas the main advantage of DEA lies in its general nonparametric treatment of the frontier. A large number of different DEA and SFA estimators have been presented during the past three decades; see Fried et al. (2008) for an up-to-date review.

In recent years, many new semi- and nonparametric stochastic frontier techniques have been developed both to relax some of the restrictive assumptions used in fully parametric frontier models and to narrow the gap between SFA and DEA. In the presence of panel data, Park et al. (1998, 2003, 2006) presented several semiparametric SFA models based on different assumptions concerning the dynamic specification of the model and joint distribution of inefficiencies and the regressors. Although the proposed semiparametric panel data models relax the assumption about inefficiency distribution, the functional form representing the production technology is still assumed to be known apart from a finite number of unknown parameters. Adams et al. (1999) further extended these approaches by developing a semiparametric panel data estimator that relaxes the distributional assumption for inefficiency and does not specify functional form for a subset of regressors. On the other hand, in a cross-sectional setting different kind of semiparametric approach was considered by Fan et al. (1996), who estimated a SFA model where the functional form of the production frontier is not specified a priori, but distributional assumptions are imposed on error components as in Aigner et al. (1977). In addition to various semiparametric SFA approaches, Kneip and Simar (1996), Henderson and Simar (2005) and Kumbhakar et al. (2007) have proposed fully nonparametric stochastic frontier techniques based on kernel regression, local linear least squares regression and local maximum likelihood, respectively. From these nonparametric approaches, the first two require panel data, while the third was developed for a cross-sectional setting.

Although the assumptions required by the aforementioned semi- and nonparametric stochastic frontier approaches are weak compared to parametric approaches, there is no guarantee that the frontiers estimated with these techniques would satisfy any regularity conditions of microeconomic theory. This is not unexpected, as these approaches were not developed to account for shape constraints such as monotonicity, concavity or homogeneity. Instead of shape constraints, the techniques used for estimating semi- or nonparametric frontier functions assume the frontier to be smooth (i.e. differentiable) and require one to specify bandwidth or other smoothing parameter prior to estimation. Nevertheless, since the smoothness assumptions are often arbitrary and the results can be very sensitive to the value of the smoothing parameter, in many applications it can be more justified to impose certain shape constraints than to specify a value for the smoothing parameter. In fact, as demonstrated by popular nonparametric DEA estimators, it is even possible to avoid smoothness assumptions completely by employing shape constraints. However, although DEA estimators can satisfy different regularity constraints by construction, they count all deviations from the frontier as inefficiency, completely ignoring all stochastic noise in the data. Due to the exclusion of noise, DEA as well as the recently developed, more robust, order-m and order-α frontier estimators, are fundamentally deterministic.[1] Hence, it is generally important to develop semi- and nonparametric approaches that are both stochastic, and similarly with DEA and some other deterministic frontier techniques, use shape constraints instead of smoothness assumptions. Besides technical efficiency measurement, these kinds of approaches are needed in environmental and economic efficiency analysis, where it is very often justified to assume that the frontier satisfies certain shape constraints.

To our knowledge, so far there have been only a few studies that have examined the estimation of semi- and nonparametric stochastic frontier models under shape constraints. Banker and Maindiratta (1992) proposed a maximum likelihood model that combines a DEA-style shape-constrained nonparametric frontier with a SFA-style stochastic composite error. However, because their model is extremely demanding computationally, it has not been estimated in any empirical applications. Kuosmanen and Kortelainen (2007) suggested

[1] For the developments in frontier estimation using deterministic approaches that are more robust to outliers and/or extreme values than DEA, see Cazals et al. (2002) and Aragon et al. (2005). In addition, Martins-Filho and Yao (2007, 2008) have recently presented two smooth nonparametric frontier estimators that are also more robust for outliers than DEA. In any event, all these estimators are deterministic in the sense that they do not separate efficiency from the statistical noise contrary to stochastic frontier estimators.

a similar kind of stochastic frontier approach, where the shape of the frontier is estimated nonparametrically using shape-constrained nonparametric least squares. They call this model as Stochastic Nonparametric Envelopment of Data (StoNED). In contrast to Banker and Maindiratta (1992), their nonparametric least squares approach is computationally feasible and can be applied quite straightforwardly, as it is based on quadratic programming.

Although the approach developed by Kuosmanen and Kortelainen (2007) can be applied for the estimation of shape-constrained stochastic frontiers in various kinds of settings, similarly to many other nonparametric methods, the precision of the shape-constrained least squares estimator decreases rapidly as the number of explanatory variables (i.e. inputs) increases. This phenomenon, known as "*the curse of dimensionality*" in nonparametric regression, implies that when data include several input variables (i.e. 3 or more), one needs very large sample size to obtain a reasonable estimation precision. This weakness of nonparametric least squares estimator is essential, because in many applications, the number of inputs is greater than 2, while the sample size is moderate. As relatively small samples with many input variables are commonly used in stochastic frontier applications, it is also important to explore flexible approaches that are not sensitive to dimensionality, but still allow one to impose shape constraints.

In this paper, our main objective is to extend the work of Kuosmanen and Kortelainen (2007) to semiparametric frontiers by developing a new approach which avoids the curse of dimensionality but allows us to impose regularity conditions on the frontier function. The shape-constrained semiparametric specification we propose is based on the single-index model, which is one of the most popular semiparametric models in econometrics literature. For the estimation of the model, we develop a three stage approach. While the first stage applies either sliced inverse regression or a monotone rank correlation estimator (both of which are common single-index estimation techniques), the second and third stages are based on similar estimation techniques used for the StoNED model. However, in contrast to StoNED estimation, our approach is not sensitive to the curse of dimensionality, because the second stage in the proposed framework is always univariate regression regardless of the number of inputs.

In addition to developing a new method for semiparametric frontier estimation, we show how the proposed approach can be modified for environmental production technology estimation in pollution generating industries. Following standard environmental economics and frontier approaches, we estimate an environmental production function by modeling emissions as inputs. In the empirical application of the paper, we illustrate the proposed semiparametric approach in environmental technology estimation with data on U.S. coal-fired electric power plants. We estimate environmental sensitive technical efficiency scores using the methods proposed in the paper and some traditional frontier methods.

The remainder of the paper is organized as follows. Section 2 presents the StoNED model and shows how it can be estimated by using shape-restricted nonparametric least squares. Section 3 proposes a shape-constrained single-index frontier model and a three stage approach for estimating the model. In Section 4 we show how the proposed approach can be modified for environmental production frontier estimation. Section 5 illustrates the developed methods using an empirical application to electric power plants. Section 6 presents the conclusions.

2. Estimation of shape-constrained nonparametric frontier

Since the semiparametric approach proposed in this paper is closely related to the StoNED approach and applies the same estimation techniques, we start by presenting the StoNED model and show how it can be estimated. For further technical details concerning this section, we refer to Kuosmanen and Kortelainen (2007) (hereafter KK).

Let us consider a multi-input single-output setting, where m-dimensional input vector is denoted by $\mathbf{x}$, the scalar output by y and deterministic production technology by the *production function* $f(\mathbf{x})$. In contrast to parametric SFA literature, we do not assume any functional form for the production function, but in the line with DEA, we require that function f belongs to the class of continuous, monotonically increasing and globally concave functions, denoted by

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

Further, we follow SFA literature (and deviate from DEA) by introducing a two-part composed error term $\varepsilon_i = v_i - u_i$, in which the second term u_i is a one-sided technical inefficiency term and the first term v_i is a two-sided statistical disturbance capturing specification and measurement errors. Using this notation, we consider the following stochastic production frontier model (or composed error model):

$$y_i = f(\mathbf{x}_i) + \varepsilon_i = f(\mathbf{x}_i) + v_i - u_i, \quad i = 1,...,n \tag{2}$$

where it is assumed that $u_i \underset{i.i.d}{\sim} \left|N(0,\sigma_u^2)\right|$, $v_i \underset{i.i.d}{\sim} N(0,\sigma_v^2)$ and that u_i and v_i $(i = 1,...,n)$ are statistically independent of each other as well as of inputs $\mathbf{x}_i$. Of course, following SFA literature, other distributions such as gamma or exponential could be used for the inefficiency term u_i (see e.g. Kumbhakar and Lovell, 2000). However, here we follow the standard practice and assume the half-normal specification.

Following KK, the model (2) is referred to *stochastic nonparametric envelopment of data* (StoNED) model. It is worth noticing that StoNED model has links to parametric SFA as well as nonparametric DEA models. Firstly, if *f* is restricted to some parametric functional form (instead of the class F_2), SFA model by Aigner et al. (1977) is obtained from (2). Secondly, if we impose the restriction $\sigma_v^2 = 0$ and relax the assumptions concerning the inefficiency term, the resulting deterministic model is similar to the single-output DEA model with an additive output-inefficiency, first considered by Afriat (1972). Thus, in contrast to other SFA models presented in literature, the StoNED model clearly connects to DEA, as monotonicity and convexity assumptions are required but no *a priori* functional form for frontier is assumed.

Standard nonparametric regression techniques cannot be used directly to estimate model (2), because $f(\mathbf{x}_i)$ is not the conditional expected value of y_i given $\mathbf{x}_i$: $E\left(y_i \middle| \mathbf{x}_i\right) = f(\mathbf{x}_i) - E\left(\varepsilon_i \middle| \mathbf{x}_i\right) \neq f(\mathbf{x}_i)$. In fact, under the half-normal specification for the inefficiency term, we know that $E(\varepsilon_i|\mathbf{x}_i) = -E\left(u_i \middle| \mathbf{x}_i\right) = -\sigma_u\sqrt{2/\pi} < 0$ (see e.g. Aigner et al., 1977). Thus, as the expected value of the composite error term is not zero, nonparametric least squares and other nonparametric regression techniques would produce biased and inconsistent estimates. However, this problem can be solved by writing the model as

$$y_i = [f(\mathbf{x}_i) - \mu] + [\varepsilon_i + \mu] = g(\mathbf{x}_i) + \eta_i, \quad i = 1,...,n, \tag{3}$$

where $\mu \equiv E(u_i|\mathbf{x}_i)$ is the expected inefficiency and $g(\mathbf{x}) \equiv f(\mathbf{x}) - \mu$ can be interpreted as an "average" production function (in contrast to the "frontier" production function *f*), and $\eta_i \equiv \varepsilon_i + \mu$ is a modified composite error term that satisfies assumption $E(\eta_i|\mathbf{x}_i) = 0$. As the modified errors η_i satisfy standard assumptions, the average production function can be estimated consistently by nonparametric regression techniques. Further, note that because μ is a fixed constant, average function *g* belongs to same functional class F_2 as *f* (i.e. it satisfies monotonicity and concavity constraints). Thus, the frontier function *f* is estimated simply by adding up the nonparametric estimate of shape-restricted average function *g* and the expected inefficiency μ.

For estimating the shape-constrained average production function KK proposed to use a convex nonparametric least squares (CNLS) technique, which minimizes least squares subject to monotonicity and concavity restrictions. It is worth emphasizing that the CNLS technique is particularly suitable for estimating model (2), because in contrast to most other nonparametric techniques it only requires monotonicity and concavity conditions (i.e. the maintained assumptions of both StoNED and DEA models), and no further smoothness assumptions (such as the degree of differentiability and the bounds of the derivatives). Based on the insight that monotonicity and concavity constraints can be written as linear inequalities by applying Afriat's theorem (Afriat, 1967, 1972), Kuosmanen (2008) proved that the following quadratic programming problem can be used for CNLS in a multiple regression setting:

$$\begin{aligned} &\min_{\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} \eta_i^2 \text{ subject to} \\ &y_i = y_i^g + \eta_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \eta_i \\ &\alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_i \;\; \forall h,i = 1,...,n \\ &\boldsymbol{\beta}_i \ge 0 \;\forall i = 1,...,n, \end{aligned} \tag{4}$$

where η_i is the modified composite error term of equation (3) and $y_i^g = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i$ is the value of average production function *g* for observation *i*. Problem (4) includes the quadratic objective function with $n(m+1)$ unknowns and n^2+n linear inequalities. The first constraint of CNLS problem (4) is interpreted as a regression equation, while the second constraint enforces concavity similarly to the Afriat inequalities and the third constraint imposes monotonicity. It is important to notice that the constant term α_i and the slope coefficients β_{ik} (k = 1,…, m) of the regression equation are observation-specific.[2] More specifically, CNLS regression (4) estimates *n* tangent hyper-planes to one unspecified production function instead of estimating one regression equation.

Although (4) provides estimates $\hat{y}_i^g$ and tangent hyperplanes for the observed points, it does not yet give an estimator for the average function *g*. For this purpose, one can take the following piecewise linear function (or representor function)

$$\hat{g}(\mathbf{x}) \equiv \min_{i \in \{1,\dots,n\}} (\hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i'\mathbf{x}), \tag{5}$$

where $\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i$ are estimated coefficients from model (4). This function is a legitimate estimator for the shape-constrained production function, as it minimizes the CNLS problem and satisfies monotonicity and concavity constraints globally (not just in observed points).[3] Basically, (5) interpolates linearly between the solutions of problem (4) giving piecewise linear function, where the number of different hyperplane segments is chosen endogeneously and is typically much lower than *n*. Because of the piecewise linear structure, estimator (5) appears to be very similar to DEA (see KK, for a graphical illustration). However, it is worth emphasizing that $\hat{g}(\mathbf{x})$ does not yet estimate the frontier, but the average production function $g(\mathbf{x})$. Nonetheless, in this framework the shape of the frontier $f(\mathbf{x})$ must be exactly the same as that of the average practice and the difference between functions results only from the expected inefficiency (compare formula (3)).

---

[2] The slope coefficients $\boldsymbol{\beta}_i$ are so-called Afriat numbers and represent the marginal products of inputs (i.e., the sub-gradients $\nabla g_i(\mathbf{x})$).

[3] Since estimator $\hat{g}(\mathbf{x})$ gives estimates also for unobserved points, it can be used, for example, to estimate substitution and scale elasticities.

To obtain estimates for production frontier and inefficiency of firms, one first needs to estimate the expected inefficiency μ and the unknown parameters σ_u, σ_v from the CNLS residuals $\hat{\eta}_i$ given by model (4). Estimation can be done straightforwardly using the method of moments (MM) which is a standard technique in stochastic frontier literature (see e.g. Kumbhakar and Lovell, 2000).[4] Having obtained estimates $\hat{\sigma}_u, \hat{\sigma}_v$ with MM, the frontier production function *f* can then be consistently estimated by

$$\hat{f}(\mathbf{x}_i) = \hat{g}(\mathbf{x}_i) + \hat{\mu} = \hat{g}(\mathbf{x}_i) + \hat{\sigma}_u \sqrt{2/\pi}\ . \tag{6}$$

Hence, similarly to the frequently used MOLS approach, production frontier is obtained by shifting the average production function upwards by the expected value of the inefficiency term.

The estimation of the technical inefficiency score for a particular observation is based on the Jondrow et al. (1982) formula:

$$E(u_i | \varepsilon_i) = \mu_* + \sigma_* \left[\frac{\phi(-\mu_* / \sigma_*)}{1 - \Phi(-\mu_* / \sigma_*)} \right], \tag{7}$$

where $\mu_* = -\varepsilon_i \sigma_u^2 / (\sigma_u^2 + \sigma_v^2)$, $\sigma_*^2 = \sigma_u^2 \sigma_v^2 / (\sigma_u^2 + \sigma_v^2)$ and $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions, respectively. The conditional expected value of inefficiency for firm *i* is calculated by substituting estimates $\hat{\sigma}_u, \hat{\sigma}_v$ and $\hat{\varepsilon}_i = \hat{\eta}_i - \hat{\sigma}_u \sqrt{2/\pi}$ in formula (7). However, as usual, this formula can only be used as a descriptive measure in a cross-sectional setting, because it is not a very good predictor for u_i.[5]

It is important to notice that the StoNED model presented above assumes an additive structure for the composite error term. This is opposite to most SFA applications that are based on the multiplicative error model

$$y_i = f(\mathbf{x}_i; \beta) \exp(v_i - u_i), \tag{8}$$

[4] Alternatively, instead of MM one could use pseudolikelihood (PSL) approach developed by Fan et al. (1996). Both MM and PSL are consistent under similar conditions, but the latter is computationally somewhat more demanding. Because of this, in this paper we apply more standard MM technique.

[5] In the cross-sectional setting Jondrow et al. formula is an unbiased but inconsistent estimator for u_i, as the variance of the estimator does not converge to zero.

which is prior to estimation transformed into the additive form by taking logarithms of both sides of equation.[6] Although both additive and multiplicative models typically assume homoskedasticity of error terms, the latter is normally less sensitive to heteroskedasticity problem than the former. This is especially true if heteroskedasticity is related to firm size, which is quite typical in applications where firms are of notably different sizes. Since the multiplicative error structure can remove or alleviate potential heteroskedasticity, in some applications it can be useful to apply StoNED with a multiplicative error structure. However, as no parametric functional form for f is specified, it is more natural to use an alternative multiplicative error model

$$y_i = \exp\left[f\left(\mathbf{x}_i\right)\right]\exp(v_i - u_i)\,, \tag{9}$$

where $f\left(.\right) \in F_2$ and error terms are assumed to have the same distribution as before. Importantly, (9) can also be transformed into additive form by taking logarithms. This implies that estimation techniques elaborated above can be applied for the model, where the dependent variable is logarithmic output and independent variables (or inputs) are expressed in levels. However, it is important to notice that in this framework shape constrains are imposed for the transformed model, not for the original multiplicative model (9). Thus, even though the estimated frontier function $\hat{f}(\mathbf{x})$ is always both monotonic and concave with respect to inputs, the estimated deterministic production technology $\hat{\mathbf{y}} = \exp\left[\hat{f}(\mathbf{x})\right]$ is assured to be monotonic, but not globally concave. This is because the exponential function preserves monotonicity, but not concavity. This property can be seen both as a weakness and strength of model (9). If one wants to impose production technology as concave with respect to inputs, this model is not sufficient for that purpose in contrast to a model with an additive error structure. On the other hand, as the multiplicative model does not require production technology to be concave, this can be a more natural framework in applications, where concavity is not a well-grounded assumption.

[6] For example, the frequently applied Cobb-Douglas and translog functional forms are based on the log-transformation of the multiplicative error model.

3. Estimation of shape-constrained single-index frontier

3.1. Background

Although StoNED models with an additive and multiplicative error structure can be estimated in various kinds of applications, there are some aspects that restrict the applicability of these approaches. One important constraint is related to the nonparametric functional form of the production function. Besides being an important strength, it can be also seen as a weakness of the StoNED approach. This is because the nonparametric function simultaneously allows great functional flexibility, but also sets considerable demands on the data set used in the application. In practice, the problem is that the precision of the nonparametric least squares estimator decreases rapidly as the number of explanatory variables (i.e. inputs) increases. This phenomenon, which is general in nonparametric regression and known as the "*curse of dimensionality*", implies that when data includes several input variables (usually 3 or more) very large sample is needed to obtain acceptable estimation precision (see e.g. Yatchew, 2003, for detailed discussion).

As relatively small samples with many input variables are commonly used in frontier applications, there is a need for shape-constrained semiparametric approaches that are not sensitive to dimensionality. Although some methods for the estimation of semiparametric stochastic frontier functions have been presented (see e.g. Fan et al., 1996; Adams et al., 1999), these techniques were not developed for estimation under regularity conditions. In addition, they assume a smooth frontier function and require one to specify bandwidth prior to estimation. Since no shape constraints are utilized, these techniques can be very sensitive to the chosen bandwidth value. Due to these deficiencies, it is important to examine the estimation of semiparametric stochastic frontier functions under shape constraints in detail.

In the next subsections we develop a shape-constrained semiparametric approach for frontier estimation based on the single-index model. It is worth noting that the presented model can be seen as the extension of the more general StoNED framework. By making stronger assumptions on the functional form than in StoNED but less restrictive than in parametric models, this model offers a compromise between StoNED and parametric shape-restricted approaches. Importantly, the proposed semiparametric approach has both advantages and weaknesses in comparison to StoNED. The main advantage is the estimation precision that can be increased by assuming a semiparametric functional form. This means that this approach can usually be applied in applications where the number of observations is small and/or there are many explanatory variables. In addition, in a multiple-input setting, the proposed estimation techniques are also computationally less demanding than the estimation approach presented in Section 2. On the other hand, it should be noted that there is always a trade-off between the estimation precision and the flexibility of the functional form specification, as additional assumptions on functional form also increase the risk of specification errors.

3.2. Single-index model

In econometric and statistics literature, various semiparametric regression models have been developed. This section presents a semiparametric model that does not suffer from the curse of dimensionality problem, and thus, allows one to include as many inputs or explanatory variables as needed in the analysis. The proposed approach is based on the single-index model (e.g. Härdle and Stoker, 1989; Ichimura, 1993), which is one of the most referred semiparametric regression models and has been widely used in various kinds of econometric applications.[7] The single-index model is based on the following specification:

$$y = g\left(h\left(\mathbf{x};\boldsymbol{\delta}\right)\right) + \varepsilon, \tag{10}$$

where $\boldsymbol{\delta}$ is a m×1 unknown parameter vector to be estimated, the function $h(.)$ (called index function) is known up to a parameter vector $\boldsymbol{\delta}$, $g(.)$ is an unknown function and ε is an unobserved random disturbance with $E\left(\varepsilon|\mathbf{x}\right)=0$. The statistical problem is to estimate the parameter vector $\boldsymbol{\delta}$ and conditional mean function g from a sample $\left\{\left(y_i,\mathbf{x}_i\right), i=1,..., n\right\}$. Note that the whole model as well as $g\left(h\left(\mathbf{x};\boldsymbol{\delta}\right)\right)$ are semiparametric, since $h\left(\mathbf{x};\boldsymbol{\delta}\right)$ is a *parametric* function and $\boldsymbol{\delta}$ lies in a finite-dimensional parameter space, while g is a nonparametric function belonging to the infinite-dimensional parameter space.

[7] See Geenens and Delecroix (2006) for the survey of the single-index model and its estimation techniques, and Yatchew (2003) for application examples.

Although it is possible to assume different kinds of functional forms for index function $h(.)$, most typically the linear index $h(\mathbf{x};\boldsymbol{\delta}) = \boldsymbol{\delta}'\mathbf{x}$ is assumed. Model (10) with $h(\mathbf{x};\boldsymbol{\delta}) = \boldsymbol{\delta}'\mathbf{x}$ is called a *linear single-index model* (e.g. Ichimura, 1993). In the context of production function and frontier estimation, use of linear single-index models implies that we assume an unknown production function to depend on a linear index of inputs, but no parametric functional form is assumed for this relationship. For simplicity, in this paper we will assume a linear index function and thus, the "single-index model" will always refer to the linear single-index model. Nevertheless, we note that in some frontier applications alternative or more general parametric functional forms (than linear) can be more appropriate for index function. It is, for example, possible to include cross products (or interactions) of explanatory variables in the index function (e.g. Cavanagh and Sherman, 1998).

It is important to notice that in single-index models some normalization restrictions are generally required to guarantee the identification of the parameter vector.[8] First of all, the matrix of explanatory variables **X** is not allowed to include a constant term. This restriction is called *location normalization*. The second restriction, called *scale normalization*, requires that one of the δ_k (k=1,...,m) coefficients is imposed to equal one.[9] This means that we can only identify the direction of the slope vector $\boldsymbol{\delta}$, that is, the collection of ratios $\{\delta_{\mathrm{j}}/\delta_{\mathrm{k}}, j,k = 1,...,m\}$, not the length or orientation of coefficients. Without lost of generality, we will thus set the first component of $\boldsymbol{\delta}$ to unity and denote the parameter vector to be estimated as $\boldsymbol{\beta}' = (1 \quad \delta_2 \quad \dots \quad \delta_m)'$. Location and scale normalization have to be imposed, because otherwise it would not be possible to uniquely identify the index function. Besides these two normalizations, it is also required that **X** includes at least one continuously distributed variable, whose coefficient is not zero and that there does not exist perfect multicollinearity between components of **X**. In addition, depending on the used estimation technique some assumptions about nonparametric function *g* are needed to avoid perfect fit.

---

[8] Identification of single-index models is discussed in detail by Ichimura (1993).
[9] There are also some other possibilities for scale normalization, see Ichimura (1993).

3.3. Estimation techniques

The main challenge in estimating single-index models is not the estimation of nonparametric function g, but the parameter vector $\boldsymbol{\beta}$. In fact, given an estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, $g(\hat{\boldsymbol{\beta}}'\mathbf{x})$ can be estimated using any standard nonparametric regression techniques (e.g. Geenens and Delecroix, 2006). However, as our aim is to develop an approach for shape-constrained production frontier estimation similarly as in Section 2, we need a technique that allows us to estimate the nonparametric function g under regularity conditions. Although it would be possible to use some other shape-constrained estimation techniques in the case of one explanatory variable (i.e. estimated single-index $\hat{\boldsymbol{\beta}}'\mathbf{x}$), analogously with the StoNED approach presented in Section 2 we will use CNLS for the estimation of average function g. By using CNLS, we do not need to assume differentiability of the frontier function or any other smoothness properties. This is in contrast to other shape-restricted nonparametric estimation techniques such as smoothing spline or Sobolev least squares (see e.g. Yatchew, 2003), which require one to specify a value for smoothing parameter in addition to shape constraints.

With regard to the estimation of single-index coefficient vector $\boldsymbol{\beta}$, there does not exist one method above the others, as various techniques have their own benefits and weaknesses. This same fact also explains why there is a great variety of methods available for single-index models. Most estimators can be classified into two main categories: the M-estimators and direct estimators. Typical examples of M-estimators include semiparametric nonlinear least squares estimator (Ichimura, 1993) and semiparametric maximum likelihood estimator (Delecroix et al., 2003), while most popular direct estimators are average derivative method (Härdle and Stoker, 1989), density-weighted average derivative estimator (Powell et al., 1989) and sliced inverse regression (Li, 1991; Duan and Li, 1991). The advantage of direct estimators is that they provide an analytic form and are therefore computationally relatively easy to implement. Instead, M-estimators have somewhat better theoretical properties, but they are also computationally much more demanding, as they require the solving of nonlinear optimization problem with nonconvex (or nonconcave) objective function. In addition to direct and M-estimators, some other estimators for index coefficients have been developed such as monotone rank correlation estimator (Cavanagh and Sherman, 1998).

In this paper, we will show how the sliced inverse regression (SIR) and the monotone rank correlation (MRC) estimator can be used for estimating the single-index coefficient vector $\boldsymbol{\beta}$ in stochastic frontier estimation.[10] As these two estimators are based on different assumptions and computational procedures, the use of both methods in a typical empirical application can make the analysis more robust. Therefore, we will also apply both techniques in the empirical application. There are two important reasons for the selection of SIR and MRC among many possibilities in this context. First of all, both techniques are based on assumptions that are consistent with the assumptions used in the second stage of our approach. In fact, to our knowledge SIR and MRC are the only single-index estimators that do not require the conditional mean function g to be differentiable. Since we use the non-smooth CNLS for estimating the nonparametric function in the second stage, here it would thus be questionable to use techniques that require the differentiability of g for the estimation of index parameters. The second relevant reason to prefer MRC and SIR to other possible estimators is related to the choice of the smoothing parameter. In contrast to all other single-index estimators mentioned above, MRC does not require bandwidth or a tuning parameter of any other kind. Instead, in SIR estimation one has to choose the number of slices, which is partially similar to bandwidth choice used in kernel regression. However, the number of slices for SIR is generally less crucial than the selection of bandwidth for typical nonparametric regression or density estimation problems (see Li, 1991, for discussion). Due to these important properties, we consider SIR and MRC the most suitable estimation techniques for the parametric part of the shape-restricted average production function.

3.4. Frontier estimation

Single-index models and techniques have been utilized in various kinds of econometric applications, including binary response, censored regression and sample selection models. Nevertheless, applications in the field of production economics have been rare, and we are aware of only two studies that have used the single-index model in production function estimation. Das and Sengupta (2004) used the single-index model to estimate both production and utilization functions for Indian blast furnaces, while Du (2004) proposed single-index specification for the deterministic frontier model that does not account for

[10] I am thankful to Leopold Simar for the suggestion to use the rank correlation estimator.

shape constraints. To avoid the dimensionality problem, the single-index model is not so advantageous in deterministic frontier estimation, since one can estimate (deterministic) nonparametric quantile frontiers in a parametric convergence rate (see Aragon et al., 2005; Martins-Filho and Yao, 2008). However, this is not the case with stochastic frontier estimation, and thus single-index model can be a much more useful tool in stochastic frontier application than in deterministic ones. Moreover, as it does not require the specification of functional form for production function *a priori*, it is important to consider how single-index specification can be used in stochastic frontier estimation in general and in shape-restricted estimation, in particular.

Let us now consider a stochastic frontier model based on the single-index specification. We assume that the frontier function f belongs to the shape-restricted class F_2 and that it has a single-index structure (10). This implies that the production frontier is monotone increasing and concave with respect to the index function. Semiparametric SFA model with an additive error structure and the same error term assumptions as before (see Section 2) can be written as

$$\begin{aligned} y_i &= f\left(\boldsymbol{\beta}'\mathbf{x}_i\right)+\varepsilon_i = f\left(\boldsymbol{\beta}'\mathbf{x}_i\right)-\mu+\varepsilon_i+\mu \\ &= g\left(\boldsymbol{\beta}'\mathbf{x}_i\right)+\eta_i, \qquad i=1,\dots,n \end{aligned} \tag{11}$$

where $\varepsilon_i = v_i - u_i$ is the composed error term, μ is the expected inefficiency, $g(.) = f(.) - \mu \in F_2$ is the average production function and $\eta_i \equiv \varepsilon_i + \mu = v_i - u_i + \mu$ is the modified composite error term with $E\left(\eta_i \middle| \mathbf{x}_i\right) = 0$. Note that the frontier function f and the average production function g have the same index functions, as constant μ only affects location, not index (which cannot have a constant). Because of this property, it is possible to estimate the single-index coefficient vector using the average production function g.

It is also important to note that the above single-index specification can easily be modified for a frontier model with a multiplicative error structure (9). This multiplicative model uses logarithmic output as dependent variable, but is otherwise similar to (11). Hence, the estimation techniques elaborated below can be also used for estimating a single-index frontier with multiplicative error structure.

For the estimation of the single-index frontier model, the following three stage procedure can be used:

[1] Estimate the coefficient vector $\boldsymbol{\beta}$ by using either sliced inverse regression (SIR) or the monotone rank correlation estimator (MRC) and calculate the values of index functions $z_i = \hat{\boldsymbol{\beta}}'\mathbf{x}_i$, i $=1,\ldots,n$ with the given estimates.

[2] Use the shape-restricted univariate CNLS (4) to estimate fitted values of the average production function $g(z_i)$. (To estimate average function for unobserved values of z, use (5).)

[3] Use the method of moments to estimate error term parameters and frontier function and Jondrow et al. measure (7) to calculate inefficiency scores.

Estimation techniques used in stages [2] and [3] have been explained in Section 2, so we skip these stages here and concentrate on stage [1]. We next describe the main principles of SIR and MRC that are used in the first stage and then comment on the statistical properties of the proposed three stage approach.

Sliced inverse regression was proposed for the purpose of dimension reduction by Li (1991). The basic principle behind the method is simple; parameter vector $\boldsymbol{\beta}$ is estimated by using inverse regression $E(\mathbf{x}|y)$, where the vector of explanatory variables **x** is explained by y. The inverse regression of **x** on y is based on a nonparametric step function as elaborated below. Computationally, SIR is probably the easiest single-index technique, because it does not require iterative computation and basically can be implemented with any econometric or statistical program. Related to this, the method is feasible and not computationally demanding to use even if the number of explanatory variables is very large.[11] On the other hand, in contrast to other single-index techniques, SIR requires an assumption that for any $\mathbf{b} \in R^m$, the conditional expectation $E(\mathbf{b}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x} = z)$ is linear in z. Li (1991) has shown that this condition can be satisfied if the matrix of explanatory variables

[11] For example, Naik and Tsai (2004) estimated a single-index model with 2424 observations and 166 explanatory variables using SIR, although only 16 of the variables proved to be significant.

X is sampled randomly from any nondegenerate elliptically symmetric distribution (such as multivariate normal distribution). This can be restrictive assumption in some applications, even though it has been shown that the linearity assumption generally holds as a reasonable approximation, when the dimension of **x** is large (see Hall and Li, 1993).

As far as the estimation procedure is concerned, SIR is quite different in comparison to most other regression techniques. In SIR, the parameter vector $\boldsymbol{\beta}$ is estimated by using the principal eigenvector $\boldsymbol{\gamma}_1$ of the spectral decomposition formula:

$$\Sigma_{\mathbf{x}|y}\,\boldsymbol{\gamma}_1 = \lambda_1\,\Sigma_{\mathbf{x}}\,\boldsymbol{\gamma}_1, \tag{12}$$

where λ_1 is the largest eigenvalue (i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$), $\Sigma_{\mathbf{x}}$ is the covariance matrix of **x**, and $\Sigma_{\mathbf{x}|y} = Cov\left(E(\mathbf{x}|y)\right)$ is the covariance matrix of the conditional mean of **x** given y. Formula (12) can be used for calculating $\boldsymbol{\beta}$ after $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}|y}$ have been substituted by their estimates. $\Sigma_{\mathbf{x}}$ can be estimated by the usual sample covariance matrix $\hat{\Sigma}_{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'$, where $\mathbf{x}_i$ denotes the values of inputs for observation *i* and $\bar{\mathbf{x}}$ contains means of input variables. Estimation of $\Sigma_{\mathbf{x}|y}$ requires that the range of output *y* is first partitioned into *Q* slices $\{s_1, \ldots, s_Q\}$, and then the *m*-dimensional conditional mean function (or inverse regression) $\boldsymbol{\xi} = E(\mathbf{x}|y)$ for each slice s_q is estimated by the sample average of the corresponding $\mathbf{x}_i$'s, that is

$$\hat{\boldsymbol{\xi}}_q = \frac{\sum_{i=1}^{n}\mathbf{x}_i 1(y_i \in s_q)}{\sum_{i=1}^{n} 1(y_i \in s_q)} \quad \text{if } y \in s_q, \tag{13}$$

where 1(.) is the indicator function taking value 1 and 0 depending on whether y_i falls into the *q*th slice or not. $\Sigma_{\mathbf{x}|y}$ can then be estimated by using a weighted sample variance-covariance matrix

$$\hat{\Sigma}_{\mathbf{x}|y} = \sum_{q=1}^{Q}\hat{p}_q\left(\hat{\boldsymbol{\xi}}_q - \bar{\mathbf{x}}\right)\left(\hat{\boldsymbol{\xi}}_q - \bar{\mathbf{x}}\right)', \tag{14}$$

where $\hat{p}_q$ is the proportion of observations in slice q. By substituting the estimates $\hat{\Sigma}_{\mathbf{x}}$ and $\hat{\Sigma}_{\mathbf{x}|y}$ into (12), we can obtain a SIR estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\gamma}}_1$ (i.e. the principal eigenvector of the spectral decomposition). Furthermore, it is then straightforward to calculate $z_i = \hat{\boldsymbol{\beta}}'\mathbf{x}_i$ for all observations and use these values in CNLS regression in the second stage.

It is worth emphasizing that the number of slices Q used in (13) and (14) has to be chosen before the estimation. However, the choice of Q does not usually affect the SIR estimates, as long as the sample size is large enough to provide useful approximations. To this end, Li (1991) showed that the number of slices for SIR is generally less crucial than the selection of bandwidth or a smoothing parameter for typical nonparametric regression or density estimation problems. In contrast to the choice of bandwidth parameter in kernel regression, the number of slices does not either affect consistency or convergence rate of the estimator (Duan and Li, 1991).

Monotone rank correlation estimator (MRC). Han (1987) first proposed an estimator based on the rank correlation between the observed dependent variable and the values fitted by the model. This maximum correlation estimator was later generalized by Cavanagh and Sherman (1998) and called a monotone rank correlation estimator (MRC). In contrast to other single-index estimators, the main benefit of MRC is that it does not require one to specify bandwidth or any other tuning parameter before the estimation. Instead, the method requires the conditional mean function g to be monotonic with respect to the index. Although this might be a restrictive assumption in certain applications, in this context it is actually very natural and justified, since we use it in stage [2].

In the single-model where the dependent variable is $\mathbf{y}$, the MRC estimator proposed by Cavanagh and Sherman (1998) uses the following objective function:

$$\hat{\boldsymbol{\beta}} = \arg\max \sum_i y_i R_n\left(\boldsymbol{\beta}'\mathbf{x}_i\right),$$

where $R_n(.)$ is the function that ranks the index values.[12] Although this may first like a relatively simple objective function, it is not easy to maximize due to the non-smooth rank

[12] For logarithmic output, one simply uses $\ln(y_i)$ in the place of y_i.

function. More importantly, since the objective function is discontinuous and thus not differentiable, it cannot be optimized with standard gradient-based algorithms (such as Newton-Raphson or BFGS). The difficulty to compute the estimator can create problems in empirical applications, since one has to rely on direct search algorithms that can locate a local optimum that is not a global optimum. In addition, search algorithms can sometimes be sensitive to the starting values of the parameters. In fact, many previous MRC studies have employed the Nelder-Mead simplex algorithm, which is not necessarily robust to starting values and the initial simplex which have to be determined before the estimation. Thus, it is possible that the simplex algorithm converges to different local maxima depending on the starting values and/or initial simplex. This potential optimization problem is demonstrated in Abrevaya (2003) who shows by means of simulations that the MRC estimator exhibit many local maxima. The results of his simulations also show that the number of local maxima increase considerably when sample size decreases. Because of these properties related to computation, at least in applications with a small sample size it might be reasonable to prefer SIR to MRC despite the weaker assumptions of the latter. On the other hand, if the used algorithm is not sensitive to the starting values or initial simplex, MRC could be more robust than the other single-index techniques, because it does not require smoothing parameter of any kind.

Asymptotic properties of estimators. Concerning the statistical properties of the proposed approach, it is worth emphasizing that the three stage method elaborated above uses estimators that are consistent under their assumptions. This means that the frontier function can also be estimated consistently if all model assumptions are valid. In addition, we have more specific asymptotic results for estimators used in different stages. First of all, $\sqrt{n}$-consistency and asymptotic normality of SIR and MRC estimators were shown by Duan and Li (1991) and Cavanagh and Sherman (1998), respectively. While SIR allows $g(.)$ to be totally unknown, its consistency depends on the linear condition explained above. Instead, the consistency of MRC is assured by the monotonicity of $g(.)$ with respect to the index. Secondly, the univariate CNLS estimator, which we use in the second stage, has been proved consistent by Hanson and Pledger (1976). Thirdly, under the stated distributional assumptions for the composed error term, error term parameters can be estimated consistently in a parametric convergence rate, even if the avarage production function is estimated with nonparametric or semiparametric methods (see Fan et al., 1996).

Besides the asymptotic results above, the benefit of the proposed approach in comparison to nonparametric frontier approaches is that it avoids the curse of dimensionality, as the frontier function can be estimated as accurately as the one-dimensional nonparametric model regardless of the number of explanatory variables. Of course, these better statistical properties can be achieved by using stronger assumptions on the structure of the model than in nonparametric estimation. Related to this, one possible weakness of the single-index model in frontier applications can be the fact that the model assumes a nonparametric functional form for the index function, not for individual variables. Despite the semiparametric treatment of the frontier, it can thus be a somewhat restrictive specification in certain applications. However, in contrast to previous techniques estimating semiparametric stochastic frontier functions, the single-index approach proposed here does not require smooth frontier and is based on shape-constrained estimation similarly to popular deterministic frontier techniques.

4. Estimation of pollution generating technology

4.1. Modelling emissions

In many industries, firms or other productions units produce undesirable outputs, such as pollution, in addition to desirable outputs. The emerging literature focuses on estimating production technologies that create pollution as a by-product of their production processes. In this literature, emissions are taken into account by estimating environmental production or frontier functions that include emissions as well as traditional inputs and outputs. We next extend the semiparametric approach proposed in the paper to the estimation of pollution generating (or environmental production) technologies. To motivate for our approach, we start by shortly reviewing various approaches used to estimate environmental production frontiers and environmentally adjusted technical efficiency or environmental efficiency scores. For brevity, we will mainly concentrate on previous SFA approaches, even though deterministic frontier approaches have been somewhat more common in the applications on this research area.

The estimation of environmental production technologies has mainly been based on DEA, deterministic parametric programming and parametric SFA methods. Evidently, the most difficult question in estimating frontier functions and/or efficiency measures in this context has been the issue of how to model emissions. In fact, although various approaches have been given justification and many academic debates have emerged, it is still open to discussion which is the "correct way" to model emissions when estimating pollution generating technologies. Following the seminal paper of Färe et al. (1989), the most common approach in DEA literature has been to model emissions as weakly disposable outputs, which basically means that the model accounts for the possibility that emissions cannot be reduced freely. However, many alternative approaches based on DEA or parametric programming have been presented and used in applications.

Instead, in classical and Bayesian SFA literature, it has been a common approach to model emissions as inputs (e.g. Koop, 1998; Reinhard et al., 1999, 2000; Managi et al., 2006). This "input approach" originates from environmental economics literature, where the standard approach of modelling nonlinear production and abatement processes is to treat waste emissions "simply as another factor of production" (Cropper and Oates, 1992). The main intuition behind this approach is that equivalently with input reduction pollution abatement is costly, as abatement requires either an increase in traditional inputs or a reduction in outputs. Therefore, it has been argued that it is justified to model emissions technically as inputs even if they represent undesirable outputs or residuals of the production in the fundamental sense. Importantly, the recent paper by Ebert and Welsch (2007) also presents a rigorous justification for the view that emissions can be modelled or interpreted as an input in the production process. In this paper, it is formally shown that a well-behaved production function with emissions as input is one of the three equivalent ways to model a production technology if the material balance is accounted for as an additional condition[13]. This result is of great importance, as some previous studies (e.g. Coelli et al., 2007) have argued conversely that the input approach is not consistent with the material balance condition.

[13] According to the material balance condition (or the law of mass conversion), the flow of materials taken from the environment for economic use, generates a flow of materials with an equal weight back into the environment.

Two notable exceptions for the input approach in SFA literature are Fernandez et al. (2002), where emissions are modelled separately from traditional inputs in a different equation, and Fernandez et al. (2005), who model emissions as normal outputs after data transformation. While the essential limitation of the former study is the separability assumption, the latter is more general in the sense that it allows nonseparability of outputs and inputs. On the other hand, to obtain a dependent variable for the regression model, Fernandez et al. (2005) need to transform emissions into desirable outputs and estimate a certain kind of parametric aggregator function that combines both untransformed and transformed outputs into one aggregated output.

4.2. Semiparametric input approach

Here we follow the standard environmental economics approach by modelling emissions as inputs in the estimation of environmental production frontiers. This means that we construct a statistical model for the good output conditional on inputs and emissions. It is worth emphasizing that treating emissions similarly to inputs simplifies estimations, as we can apply the framework proposed in Section 3. Furthermore, since the econometric estimation of multiple input, multiple output technologies is plagued with difficulties even in a fully parametric context (compare e.g. Fernadez et al., 2005), it seems sensible to use the input approach in the semiparametric estimation.

To present the idea formally, let us now denote the p-dimensional vector of emissions by **w** and the m-dimensional traditional input vector by **x**. We will now consider the function $f(\mathbf{x},\mathbf{w})$, which we call environmental production frontier. By following Section 3, we will assume that this function takes the single-index form $f(\mathbf{x},\mathbf{w}) = f(\boldsymbol{\beta}'\mathbf{x}+\boldsymbol{\gamma}'\mathbf{w})$, where f is a nonparametric function belonging to the shape-restricted class F_2, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors and $\boldsymbol{\beta}'\mathbf{x}+\boldsymbol{\gamma}'\mathbf{w}$ is the (linear) index function. Shape constraints imply that the environmental production frontier is monotonically increasing and concave with respect to the index function.

As we model emissions similarly to traditional inputs, we can now present the frontier model simply as

$$y_i = f\left(\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}'\mathbf{w}_i\right) + \varepsilon_i = g\left(\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}'\mathbf{w}_i\right) + \eta_i, \qquad i = 1,...,n \tag{15}$$

where $\varepsilon_i = v_i - u_i$ is the composed error term, μ is the expected inefficiency, $\eta_i \equiv \varepsilon_i + \mu$ is the modified composite error term with $E\left(\eta_i | x_i\right) = 0$ and $g\left(.\right) = f\left(.\right) - \mu \in F_2$ is the average environmental production function.

To estimate model (15), we can use the three stage approach elaborated in Section 3. Having estimated residuals $\hat{\eta}_i$, environmentally adjusted technical efficiency scores (or environmental efficiency scores) can be calculated by employing the Jondrow et al. (1982) measure (7). There are some other measures available for the environmental efficiency estimation in SFA literature, but these require a parametric functional form for the environmental production function (see Reinhard et al., 1999; Fernandez et al., 2005).

5. Application to electric power plants

5.1. Data and estimations

In this section the proposed semiparametric techniques are applied to empirical data both to illustrate the new techniques and to compare the efficiency estimates given by these methods with those obtained by StoNED and standard DEA and SFA methods. We estimate an environmental production frontier and environmentally adjusted technical efficiency scores for a set of U.S. coal-fired power plants by using the same data set as in Färe et al. (2007a). This data set includes 92 observations from year 1995 and is based on the larger database used by Pasurka (2006) and Färe et al. (2007b). It is important to notice that these data only include plants in which at least 95% of total fuel consumption (in Btu) is provided by coal. This guarantees that the plants included in the data are comparable with respect to their production technology.

For estimating the frontier functions and efficiency scores, we will use one desirable output, two different emissions and two inputs. The desirable output is net electrical generation in gigawatt-hours (GWh) and pollution variables include sulfur dioxide (SO_2) and nitrogen oxides (NO_x) emissions. Input variables consist of capital stock measured in 1973 million dollars and the annual average number of employees at the plant. Concerning the data sources, net electrical generation and fuel consumption data come from the *Annual Steam Electric Unit Operation and Design Report*, published within the Department of Energy (DOE) by the Energy Information Administration, EIA767. These data are also used by DOE to derive emission estimates of SO_2 and NO_x. Capital and labor data is based on the information compiled by the US Federal Energy Regulatory Commission (FERC). For details on how the data set has been constructed and on different assumptions made to elaborate the variables, we refer to Färe et al. (2007a). Table 1 presents descriptive statistics for each variable used in the analysis.

Table 1. Descriptive statistics for the model variables

Variable	Units	Mean	St. dev.	Min.	Max.
Electricity	GWh	4686.5	4065.3	166.6	18212.1
Capital stock	Dollars (in millions, 1973$)	240.0	146.4	39.4	750.0
Employees	The number of workers	185.2	110.9	38.0	535.0
SO_2	Short tons	40745.2	48244.8	1293.2	252344.6
NO_x	Short tons	17494.0	16190.1	423.1	72524.1

Besides the variables in Table 1, Färe et al. (2007a) also included the heat content (in Btu) of coal, oil and gas consumed at the plant as variables in their DEA models. However, as we next argue, there are some important reasons for why these variables are not so useful in stochastic frontier estimation. First of all, it was observed in preliminary estimations that the heat content of oil and gas did not have any explanatory power for electricity in these coal-fired plants. In contrast, the heat content of coal turned out to correlate almost perfectly with electricity generation, as the correlation coefficient was as high as 0.996. Since there is a close to linear relationship between coal input and electricity, all regression models that include the heat content of coal as an explanatory variable would yield an almost perfect regression fit independently of other variables and functional form of the model. In stochastic frontier estimation, this would imply that the frontier function and average production function are equal or that there is no inefficiency according to the estimated model. This was observed in the linear and log-linear SFA models where the heat content of coal was the only explanatory variable as well as in more general models that included many input variables. However, it needs to be emphasized that this does not imply there to be no inefficiency in the utilization of some other inputs or emissions generated by the plant. Due to these reasons, we think it is justified not to include the heat content variables in the stochastic frontier models in this case.[14] For comparability, we will also exclude these variables from the DEA model we estimate. Nevertheless, we note that in DEA models the inclusion of the coal variable does not create similar problems as in SFA and there is inefficiency even after including it as the additional model variable. The reason for the divergence of DEA and SFA in these kinds of cases is left for future research.

We estimate the frontier functions and efficiency scores by using a Cobb-Douglas SFA estimator, StoNED, single-index stochastic frontier estimators based on SIR and MRC as well as a variable returns to scale (VRS) DEA estimator. Since the data include plants that differ notably with respect to their size, regression models with an additive error structure are more sensitive to the heteroskedasticity problem than models with a multiplicative error structure. As a result of this data property, we decided to use the multiplicative error specification (9) in the single-index and StoNED models. Thus, in these models the dependent variable is ln(GWh) while independent variables are measured in levels. Instead, in the variable returns to scale DEA model we use the level variable (i.e. GWh) as an output, because DEA applications based on logarithmic variables are very rare.

The DEA and parametric SFA models were estimated with Limdep. To estimate CNLS regression used in the second stage of the single-index frontier models and in StoNED, the GAMS code of Kuosmanen (2008) was used. The first stage of MRC and SIR frontier models were estimated with GAUSS and R, respectively. For the former, we employed the GAUSS code written by Jason Abrevaya, whereas the latter is based on the *dr* package in R.

As explained in Section 3, the MRC estimation requires one to use a non-gradient search algorithm to optimize the non-smooth objective function. Similarly to other previous MRC applications, an iterative Nelder-Mead simplex method was used for that purpose. For computations, we used the same iteration scheme as in Cavanagh and Sherman (1998). As starting values, we tried least squares estimates as well as some other values. Unfortunately, the coefficients were sensitive both to the starting values and the chosen initial simplex. Taking into account the simulation results of Abrevaya (2003), we doubt this problem is a

[14] However, if one would be interested in analysing the effect of various inputs on electricity generation, then it could be warranted to include the heat content of coal in the regression model.

consequence of the small sample size used in the application.[15] On the other hand, it should be noted that although the parameter estimates were sensitive to the starting values, the effect was substantially slighter on the estimated index functions. In the next section, we will give the results of the single-index model based on OLS starting values. However, because of computational problems, it is important to be cautious when interpreting the results of the MRC estimation.

In contrast to MRC, the SIR estimates are not sensitive to computational issues. However, in SIR, one has to determine the number of slices Q used in the nonparametric step function. We calculated parameter estimates and index functions with different values for Q. Although the choice of Q affected the values of coefficients, the index function estimates were very similar independently of the number of slices. As an evidence of this, the correlation coefficients between the index function estimates based on different values of Q are of important note. For example, the correlation coefficients between index functions based on Q = 3, 7 and 15 were 0.993, 0.995 and 0.9998, respectively. In the following, we report the results based on Q = 7.

5.2. Results

We start by illustrating the estimated single-index frontiers based on these data. Figures 1 and 2 plot the values of the index function (×) and single-index frontiers based on SIR and MRC. In both figures, the dependent variable ln(GWh) is on the y-axis and the index function on the x-axis. Nonetheless, the index values vary between the figures, since they are based on different methods. In both cases, the frontier functions are piecewise linear, monotonic and concave similarly to StoNED and DEA. This is because we have used CNLS in the second stage. Note that there are observations (or index values) above the estimated frontier functions. This is expected, as frontiers presented in the figures do no account for observation-specific noise terms. However, as usual in SFA, noise terms are accounted for in the estimation of inefficiency scores.

[15] Cavanagh and Sherman (1998) report that their results were not sensitive to the starting values and initial simplex. However, their sample included 18967 observations.

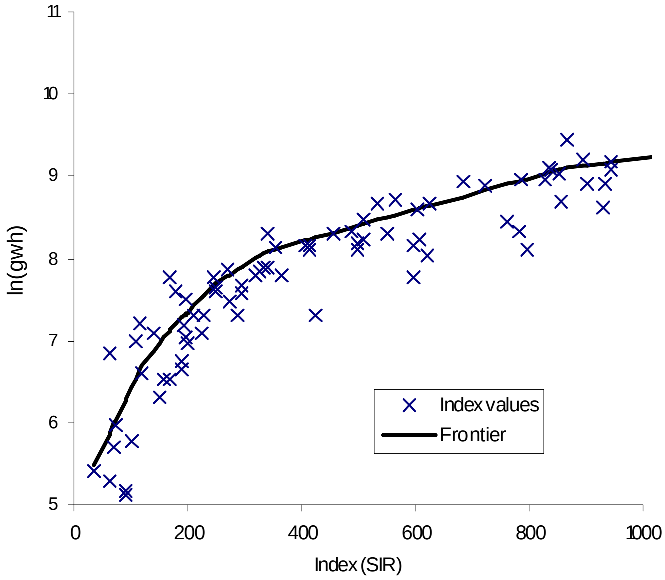

Figure 1. Single-index frontier function for SIR

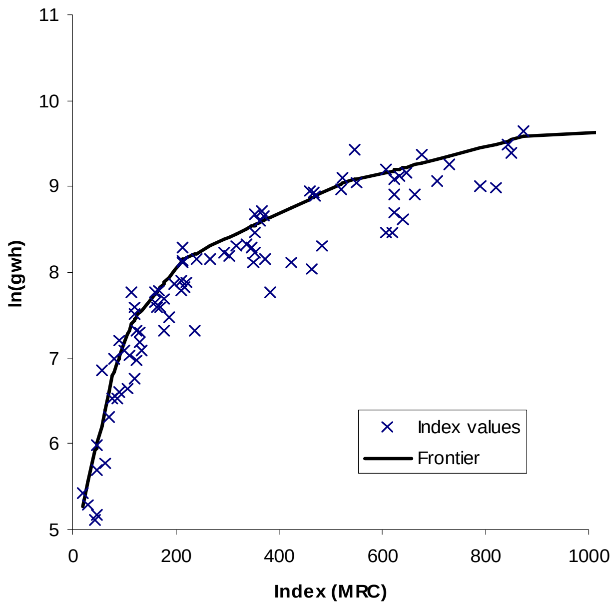

Figure 2. Single-index frontier function for MRC

Since the estimated models include four input or explanatory variables, we cannot present the estimated frontiers for StoNED, parametric SFA or DEA in figures. For the purpose of comparison, we present summary statistics of the environmentally adjusted technical efficiency scores from different models in Table 2, while Table 3 shows the correlation coefficients of efficiency scores between the methods. In addition, the appendix includes estimation results for error term parameter estimates from different stochastic frontier models. Concerning the technical efficiency scores, for all stochastic frontier models inefficiency scores were first estimated by employing the Jondrow et al. measure. Then these inefficiency scores were transformed into relative (or Farrell) efficiency scores by applying the usual formula $TE_i = \exp\left[-E(u_i|\varepsilon_i)\right]$, where $E(u_i|\varepsilon_i)$ is the Jondrow et al. measure.

Table 2. Summary statistics on environmentally adjusted technical efficiency scores

	Mean	St. dev.	Min.	Max.
Single-index, SIR	0.920	0.072	0.689	1
Single-index, MRC	0.873	0.109	0.616	1
StoNED	0.881	0.105	0.587	1
DEA (VRS)	0.737	0.207	0.273	1
SFA (Cobb-Douglas)	0.718	0.148	0.445	0.949

Table 3. Correlations of efficiency measures

	Single-index, SIR	Single-index, MRC	StoNED	DEA (VRS)	SFA (Cobb-Douglas)
Single-index, SIR	1				
Single-index, MRC	0.903	1			
StoNED	0.649	0.809	1		
DEA (VRS)	0.676	0.806	0.848	1	
SFA (Cobb-Douglas)	0.856	0.938	0.785	0.814	1

According to the results, average efficiency is highest for the SIR model and lowest for Cobb-Douglas SFA. The difference between single-index models and StoNED in average efficiency is small, whereas deviation from DEA and parametric SFA is greater. Note that the minimum value of the efficiency score is notably lower for DEA than other models. In addition, the standard deviation of efficiency scores for DEA diverges from the others.

As far as correlation of efficiency scores between the methods is concerned, the highest correlation coefficient 0,938 is between Cobb-Douglas and MRC, while the lowest is between SIR and StoNED. However, since all the correlation coefficients are yet quite high, it would be risky to present any general conclusions about the differences among the methods. Naturally, a more systematic comparison of the different techniques in small samples would require the use of simulated data sets. Nevertheless, this application demonstrates that the proposed semiparametric estimation techniques can yield empirical results that deviate from the results given by traditional DEA and SFA methods.

6. Conclusions

We have presented a new semiparametric approach for stochastic frontier estimation. We showed how the proposed shape-constrained model can be estimated in three stages by using (1) single-index estimation techniques, (2) convex nonparametric least squares (CNLS) and (3) method of moments. Importantly, as our procedure in the second and third stages is similar to the StoNED approach presented by Kuosmanen and Kortelainen (2007), the proposed approach can be considered a semiparametric extension of StoNED. Furthermore, since the second stage in our approach is always univariate regression that uses an index function as the only regressor regardless of the number of original explanatory variables, one can perceive the first stage as a dimension reduction for the second stage. This dimension reduction aspect also explains why the proposed method is not sensitive to the curse of dimensionality problem in contrast to StoNED and many other non- and semiparametric SFA approaches.

For the first stage estimation, we proposed two different methods: sliced inverse regression (SIR) and the monotone rank correlation estimator (MRC). Although there exist many other single-index estimation techniques, we considered SIR and MRC most suitable for the present model, because in contrast to all other techniques, these do not require the differentiability of the frontier function. The main benefit of MRC is that it does not need any kind of bandwidth or a smoothing parameter, which means that its estimates are not sensitive to an arbitrary smoothness assumption that is the case with most other single-index techniques. However, since the MRC estimator is based on the maximization of the non-smooth objective function, the direct search algorithm used for the estimation can be sensitive to the initial parameter values. This computational shortcoming can be especially problematic if sample size is relatively small, which was also the case in our empirical application. In contrast to MRC, SIR is generally very easy to calculate and can be implemented without any iteration procedures. However, its main weakness is the assumption on the linearity of the conditional expectation $E\left(\mathbf{b}'\mathbf{x}\middle|\boldsymbol{\beta}'\mathbf{x}=z\right)$. In addition, before estimation SIR requires one to specify the number of slices, which can have some effect on the results. All in all, since SIR and MRC have their own strengths and weaknesses, we find a good strategy to use both techniques in empirical applications. However, with access to a relatively large sample, one might prefer MRC due to its weaker assumptions.

In addition to showing how to estimate frontier and technical efficiency scores, we modified the proposed semiparametric approach for the estimation of environmental production technologies and environmental sensitive technical efficiency scores. For this purpose, we followed the standard environmental economics approach by modelling emissions as inputs. We illustrated the presented approach with an empirical application to the environmentally adjusted performance evaluation of electric power plants. Presumably due to a small sample size (n = 92), the MRC estimates were somewhat sensitive to the starting values of the used Nelder-Mead simplex algorithm. As index function estimates given by the SIR estimator were not sensitive to the number of slices, we rely more on the results given by the latter method in this application. It is left for further research to establish whether some other optimization method (or a combination of optimizers) would be more robust in MRC estimation with smaller sample sizes.

In the future, it would also be interesting and important to compare the performance of our semiparametric single-index approaches based on SIR and MRC to StoNED by employing simulated data sets. This would perhaps reveal in what kinds of settings the single-index approach is an adequate modeling tool and even preferable to StoNED. Another important research question would be to extend the proposed non-smooth approach to the estimation of smooth shape-constrained semiparametric frontier functions. In addition, it would be important to use the approaches proposed in the paper in other kinds of applications. For example, profit frontier estimation would be a natural application area, since profit functions have to satisfy shape-constraints implied by microeconomic theory.

Acknowledgements

I am grateful to Carl Pasurka for providing me with the database on U.S. coal-fired electric plants employed in the empirical application. I thank Timo Kuosmanen and Mika Linden for insightful suggestions and help with the computations carried out for the paper. Earlier versions of this paper have been presented at the 10th European Workshop on Efficiency and Productivity Analysis (EWEPA X), Lille, France, June 2007, FDPE Econometrics and Computational Economics Workshop, December 2007, and XXX Annual Meeting of Finnish Society of Economic Research, February 2008. I thank the participants of these workshops, and in particular, Heikki Kauppi, Carlos Martins-Filho, Leopold Simar and Timo Sipiläinen for their useful comments and stimulating discussion. Financial support from the Finnish Doctoral Programme in Economics (FDPE) is gratefully acknowledged.

References

Abrevaya, J. (2003): Pairwise-Difference Rank Estimation of the Transformation Model. *Journal of Business and Economics Statistics* 21(3), 437-447.

Adams, R., A. Berger and R. Sickles (1999): Semiparametric Approaches to Stochastic Panel Frontiers with Applications in the Banking Industry. *Journal of Business and Economic Statistics* 17(3), 349-358.

Afriat, S.N. (1967): The Construction of a Utility Function from Expenditure Data. *International Economic Review* 8(1), 67-77.

Afriat, S.N. (1972): Efficiency Estimation of Production Functions. *International Economic Review* 13(3), 568-598.

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977): Formulation and Estimation of Stochastic Frontier Models. *Journal of Econometrics* 6(1), 21-37.

Aragon, Y., A. Daouia and C. Thomas-Agnan (2005): Nonparametric Frontier Estimation: A Conditional Quantile-Based Approach. *Econometric Theory* 21(2), 358-389.

Banker, R.D. and A. Maindiratta (1992): Maximum Likelihood Estimation of Monotone and Concave Production Frontiers. *Journal of Productivity Analysis* 3(4), 401-415.

Cavanagh, C. and R.P. Sherman (1998): Rank Estimators for Monotonic Index Models. *Journal of Econometrics* 84(2), 351-381.

Cazals, C., J.P. Florens and L. Simar (2002): Nonparametric Frontier Estimation: A Robust Approach. *Journal of Econometrics* 106(1), 1-25.

Charnes, A., W.W. Cooper and E. Rhodes (1978): Measuring the Inefficiency of Decision Making Units. *European Journal of Operational Research* 2(6), 429-444.

Coelli, T.J., L. Lauwers and G. Van Huylenbroeck (2007): Environmental Efficiency Measurement and the Materials Balance Condition. *Journal of Productivity Analysis* 28(1-2), 3-12.

Cropper, M.L. and W.E. Oates (1992): Environmental Economics: A Survey. *Journal of Economic Literature* 30(2), 675-740.

Das, S. and R. Sengupta (2004): Projection Pursuit Regression and Disaggregate Productivity Effects: The Case of the Indian Blast Furnaces. *Journal of Applied Econometrics* 19(3), 397-418.

Delecroix, M., W. Härdle and M. Hristache (2003): Efficient Estimation in Conditional Single-Index Regression. *Journal of Multivariate Analysis* 86(2), 213-226.

Du, S. (2004): Nonparametric and Semi-Parametric Estimation of Efficient Frontier. Available at SSRN: http://ssrn.com/abstract=783827.

Duan, N. and K.C. Li (1991): Slicing Regression: A Link Free Regression Method. *Annals of Statistics* 19(2), 505-530.

Ebert, U. and H. Welsch (2007): Environmental Emissions and Production Economics: Implications of the Materials Balance. *American Journal of Agricultural Economics* 89(2), 287-293.

Fan, Y., Q. Li and A. Weersink (1996): Semiparametric Estimation of Stochastic Production Frontier Models. *Journal of Business and Economic Statistics* 14(4), 460-468.

Färe, R., S. Grosskopf, C.A.K. Lovell and C.A. Pasurka Jr. (1989): Multilateral Productivity Comparisons When Some Outputs are Undesirable: A Non-Parametric Approach. *Review of Economics and Statistics* 71(1), 90-98.

Färe, R., S. Grosskopf and C.A. Pasurka Jr. (2007a). Environmental Production Functions and Environmental Directional Distance Functions. *Energy - The International Journal* 32(7), 1055-1066.

Färe, R., S. Grosskopf and C.A. Pasurka Jr. (2007b): Pollution Abatement Activities and Traditional Measures of Productivity: A Joint Production Perspective. *Ecological Economics* 62(3-4), 673-682.

Farrell, M.J. (1957): The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society*, *Series A. General* 120(3): 253-282.

Fernandez, C., G. Koop and M.F.J. Steel (2002): Multiple-Output Production with Undesirable Outputs: An Application to Nitrogen Surplus in Agriculture. *Journal of the American Statistical Association* 97(458), 432–442.

Fernandez, C., G. Koop and M.F.J. Steel (2005): Alternative Efficiency Measures for Multiple-Output Production. *Journal of Econometrics* 126(2), 411-444.

Fried, H., C.A.K. Lovell and S. Schmidt (2008): *The Measurement of Productive Efficiency and Productivity Change*. Oxford University Press, New York.

Geenens, G. and M. Delecroix (2006): A Survey About Single-Index Model Theory. *International Journal of Statistics and Systems* 1(2), 203-230.

Hall, P. and K.C. Li (1993): On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *Annals of Statistics* 21(2), 867–889.

Han, A.K. (1987): Nonparametric Analysis of a Generalized Regression Model. *Journal of Econometrics* 35(2-3), 303-316.

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

Härdle, W. and T.M. Stoker (1989): Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of American Statistical Association* 87(417), 218-226.

Henderson, D.J. and L. Simar (2005): A Fully Nonparametric Stochastic Frontier Model for Panel Data. Discussion Paper 0417, Institut de Statistique, Universite Catholique de Louvain.

Ichimura, H. (1993): Semiparametric Least squares (SLS) and Weighted SLS Estimation of Single-Index Models. *Journal of Econometrics* 58(1-2), 71-120.

Jondrow, J., C.A.K. Lovell, I.S. Materov and P. Schmidt (1982): On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. *Journal of Econometrics* 19(2-3), 233-238.

Kneip, A. and L. Simar (1996): A General Framework for Frontier Estimation with Panel Data. *Journal of Productivity Analysis* 7(2-3), 187-212.

Koop, G. (1998): Carbon Dioxide Emissions and Economic Growth: A Structural Approach. *Journal of Applied Statistics* 25(4), 489–515.

Kumbhakar, S.C. and C.A.K. Lovell (2000): *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Kumbhakar, S.C., B.U. Park, L. Simar and E.G. Tsionas (2007): Nonparametric Stochastic Frontiers: A Local Maximum Likelihood Approach. *Journal of Econometrics* 137(1), 1-27.

Kuosmanen, T. (2008): Representation Theorem for Convex Nonparametric Least Squares. *Econometrics Journal*, forthcoming.

Kuosmanen, T. and M. Kortelainen (2007): Stochastic Nonparametric Envelopment of Data: Cross-Sectional Frontier Estimation Subject to Shape Constraints. Discussion Paper N:o 46, Economics and Business Administration, University of Joensuu.

Li, K.C. (1991): Sliced Inverse Regression for Dimension Reduction. *Journal of American Statistical Association* 86(414), 316-342.

Managi, S., J.J. Opaluch, D. Jin and T.A. Grigalunas (2006): Stochastic Frontier Analysis of Total Factor Productivity in the Offshore Oil and Gas Industry. *Ecological Economics* 60(1), 204-215.

Martins-Filho, C. and F. Yao (2007): Nonparametric Frontier Estimation via Local Linear Regression. *Journal of Econometrics* 141(1), 283-319.

Martins-Filho, C. and F. Yao (2008): A Smooth Nonparametric Conditional Quantile Frontier Estimator. *Journal of Econometrics* 143(2), 317-333.

Meeusen, W. and J. van den Broeck (1977): Efficiency Estimation from Cobb-Douglas Production Function with Composed Error. *International Economic Review* 8(2), 435-444.

Naik, P.A. and C.-L. Tsai (2004): Isotonic Single-Index Model for High-Dimensional Database Marketing. *Computational Statistics & Data Analysis* 47(4), 775 – 790.

Park, B., R.C. Sickles and L. Simar (1998): Stochastic Panel Frontiers: A Semiparametric Approach. *Journal of Econometrics* 84(2), 273-301.

Park, B., R.C. Sickles and L. Simar (2003): Semiparametric Efficient Estimation of AR(1) Panel Data Models. *Journal of Econometrics* 117(2), 279-309.

Park, B., R.C. Sickles and L. Simar (2006): Semiparametric Efficient Estimation of Dynamic Panel Data Models. *Journal of Econometrics* 136(1), 281-301.

Pasurka Jr., C.A. (2006): Decomposing Electric Power Plant Emissions within a Joint Production Framework. *Energy Economics* 28(1), 26-43.

Powell, J.L., J.H. Stock and T.M. Stoker (1989): Semiparametric Estimation of Index Coefficients. *Econometrica* 57(6),1403-1430.

Reinhard, S., C.A.K. Lovell and G. Thijssen (1999): Econometric Application of Technical and Environmental Efficiency: An Application to Dutch Dairy Farms. *American Journal of Agricultural Economics* 81(1), 44–60.

Reinhard, S., C.A.K. Lovell and G.J. Thijssen (2000): Environmental Efficiency With Multiple Environmentally Detrimental Variables; Estimated with SFA and DEA. *European Journal of Operational Research* 121(2), 287-303.

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

Appendix

Table A1. Estimates for error term parameters

	Single-index, SIR	Single-index, MRC	StoNED	SFA, Cobb-Douglas
σ_u^2	0,228	0,273	0,230	0,451
σ_v^2	0,331	0,251	0,167	0,144
σ^2	0,161	0,138	0,081	0,225
λ	0,689	1,088	1,384	3,130