mathematics

MDPI

Article

Combining Data Envelopment Analysis and Machine Learning

Nadia M. Guerrero, Juan Aparicio * and Daniel Valero-Carreras

Center of Operations Research (CIO), Miguel Hernandez University of Elche (UMH), 03202 Elche, Spain; nguerrero@umh.es (N.M.G.); dvalero@umh.es (D.V.-C.)

* Correspondence: j.aparicio@umh.es; Tel.: +34-966658517; Fax: +34-966658715

Abstract: Data Envelopment Analysis (DEA) is one of the most used non-parametric techniques for technical efficiency assessment. DEA is exclusively concerned about the minimization of the empirical error, satisfying, at the same time, some shape constraints (convexity and free disposability). Unfortunately, by construction, DEA is a descriptive methodology that is not concerned about preventing overfitting. In this paper, we introduce a new methodology that allows for estimating polyhedral technologies following the Structural Risk Minimization (SRM) principle. This technique is called Data Envelopment Analysis-based Machines (DEAM). Given that the new method controls the generalization error of the model, the corresponding estimate of the technology does not suffer from overfitting. Moreover, the notion of ε-insensitivity is also introduced, generating a new and more robust definition of technical efficiency. Additionally, we show that DEAM can be seen as a machine learning-type extension of DEA, satisfying the same microeconomic postulates except for minimal extrapolation. Finally, the performance of DEAM is evaluated through simulations. We conclude that the frontier estimator derived from DEAM is better than that associated with DEA. The bias and mean squared error obtained for DEAM are smaller in all the scenarios analyzed, regardless of the number of variables and DMUs.

Keywords: data envelopment analysis; PAC learning; support vector regression; machine learning; structural risk minimization

MSC: 90C08

Citation: Guerrero, N.M.; Aparicio, J.; Valero-Carreras, D. Combining Data Envelopment Analysis and Machine Learning. *Mathematics* **2022**, *10*, 909. https://doi.org/10.3390/math10060909

Academic Editors: Kuo-Ping Lin, Chien-Chih Wang, Chieh-Liang Wu and Liang Dong

Received: 20 February 2022
Accepted: 9 March 2022
Published: 11 March 2022

1. Introduction

One of the most important issues in the field of statistical learning is the reliability of statistical inference methods. In this framework, a sophisticated theory, the so-called Generalization Theory, explains which factors must be controlled to achieve good generalization. Optimal generalization is achieved when the error generated on evaluating new data through an inference learning method is minimized. The Generalization Theory copes with those factors that allow for the minimization of the prediction or generalization error.

In terms of pattern classifiers, the generalization error is the probability of misclassifying a randomly chosen example that holds with high probability over randomly chosen training sets, and then, a good generalization is achieved when this is minimized. This aim is possible if an upper bound of the generalization error is found, and the parameters on which it depends are controlled in order to reduce it. These bounds are understood as Probably Approximately Correct (PAC) bounds, which specifically means that the probability of the bound failing is small (Probably) when the bound is achieved through the classifier that has a low error rate (Approximately Correct). The standard PAC learning model implements the idea of finding this classifier: it considers a fixed hypothesis (classifier) class together with a required accuracy and confidence, and takes into account the theory that characterizes when a function from this class can be learned from examples (training sample) in terms of a measure called the Vapnik–Chervonenkis dimension (VC dimension).

However, the statistical learning theory (Vapnik [1]) reveals that it is much more interesting not to preselect the class that will contain the target function to be learned. Instead, it is defined a set of hypothesis classes saved as a hierarchy, and the target function to be learned lies in one of them. The Structural Risk Minimization (SRM) copes with the problem of minimizing an upper bound on the expected risk over each of these hypothesis classes (Vapnik [2]). To implement the SRM in Support Vector Machines (SVM), one must consider the structures (classes) that control two factors that appear in the bound of the expected risk: the value of empirical risk and the complexity (the appropriate bound for the generalization error). Thus, under this principle, to select a learning algorithm, it is necessary to have the theoretical bound of the generalization error (PAC bounds) and to deal with the minimization of this bound together with the empirical risk.

In addition, standard regression methods are only concerned with the minimization of the empirical risk. This is the one based on the error produced by the regressors with respect to the observed dataset. This error is defined as the distance between the data to the approximation function; thus, it is a measure of the deviation of the data with respect to the regressors. It is characterized as a residual. The vertical distance is the most common way to measure the regression error, although it is not induced by a mathematical norm. In Support Vector Regression (SVR) (Vapnik [1]), for example, the residuals that participate in the empirical risk are measured through the vertical distance. Other distances, in this case based on a norm, have been used in order to establish these residuals, such as the l_1-*distance*, the l_2-*distance* and the l_∞-*distance* (Blanco et al. [3,4]).

The estimation of production functions and measures of efficiency and productivity have been the focus of a relatively large body of articles in the literature in both the economic and engineering contexts, as well as in operations research and statistics. In particular, Data Envelopment Analysis (DEA) (Charnes et al. [5] and Banker et al. [6]) is one of the existing techniques for estimating production functions and measuring efficiency. DEA relies on the construction of a polyhedral technology in the space of inputs and outputs that satisfies certain classical axioms of production theory (e.g., monotonicity and convexity). It is a non-parametric data-driven approach with many advantages from a benchmarking point of view. Additionally, the treatment of the multi-output multi-input framework is relatively straightforward with DEA, in comparison with other methods available. However, Data Envelopment Analysis has been criticized for its non-statistical nature, even being labeled as a pure descriptive tool of the data sample at a frontier level with little inferential power (its inferential power is exclusively based on the property of consistency and the increase in sample size instead of on the fundamentals of the method) (Esteve et al. [7]). DEA suffers from an overfitting problem because of the application of the minimal extrapolation principle, which places the estimator of the production function as close to the dataset as possible. This principle is also related to exclusively minimizing the empirical error (at a frontier level).

Regarding the literature related to this topic, some previous authors have tried to modify the standard DEA technique such that the new approaches work as inferential methods (with the focus on the DGP) rather than as mere descriptive tools. For example, Banker and Maindiratta [8] and Banker [9] associated DEA with maximum likelihood. Simar and Wilson [10–12] adapted bootstrapping to DEA. Kuosmanen and Johnson [13,14] introduced the Corrected Concave Nonparametric Least Squares. Unfortunately, despite the importance of machine learning techniques in the current literature, there have been few attempts to adapt DEA to the field of machine learning (see, for example, Esteve et al. [7], or Olesen and Ruggiero [15]). In this sense, our contribution could be seen as a new bridge between these two worlds: machine learning and efficiency measurement.

In this paper, our main objective is to propose, for the first time in the literature, a PAC bound in the context of the estimation of polyhedral technologies in microeconomics and engineering, enabling the possibility of controlling the generalization error of the estimation of the production frontier. Accordingly, we construct a model that controls the empirical error, together with the generalization error, through a PAC bound implementing

the philosophy of Structural Risk Minimization by analogy with SVM. Our modeling has several implications:

(a) For the first time, a bound of the generalization error is implemented to determine the degree of technical inefficiency of a set of Decision Making Units (DMUs).
(b) We implement the minimization of the balance between the generalization error and the empirical error through a quadratic optimization model that will be called Data Envelopment Analysis-based Machines (DEAM), which has DEA as a particular case.
(c) Through a computational simulation experience, we show that DEAM outperform DEA regarding bias and mean squared error.
(d) We estimate production technologies using robust regression models that use the concept of margin. Due to that, the problem of efficiency measurement becomes a classification problem: to be efficient (being located within the margin) or not to be efficient (being located out of the margin).

Finally, we mention that the expected new insights gained by applying our approach (DEAM) are related to the determination of better estimates of production functions in engineering and microeconomics, in terms of bias and mean squared error. Additionally, these gains will also benefit the technical efficiency measures that can be derived from calculating the distance from a given observation to the production function estimate.

The rest of the paper is organized as follows. The following section provides the basic background. Next, in the third section, we introduce a new PAC for the class of piece-wise linear functions. In Section 4, a new approach called Data Envelopment Analysis-based Machines (DEAM) is defined and analyzed. Section 5 shows the main results associated with a computational experience for checking the new approach in comparison with DEA. Section 6 contains a discussion on the main results. Finally, the article ends with the conclusions section.

2. Background

In this section, we briefly introduce elemental notions of Support Vector Regression, Statistical Learning and Data Envelopment Analysis.

2.1. Support Vector Regression (SVR)

Machine learning (ML) is a methodology that studies computer processes that learn from experience and make improvements automatically. ML works with computer algorithms based on a learning sample (training data) and can make predictions about the behavior of future data. The study of this behavior is produced in two different scenarios: the scenario of supervised learning in which training data are vectors of predictors and responses, and the scenario of unsupervised learning, where no responses are considered in the data sample. In the first field, the objective of learning techniques is to determine the functional relationship between the predictors and the responses. In this case, the nature of the responses, if they come from a binary variable or are real values, determines the kind of problem to solve: a classification problem or a regression problem, respectively. In the second field, since there are no responses, the objective is to gain knowledge about the processes lying behind data generation, such as density estimation or clustering. Our paper largely focuses on the regression problem within supervised learning, bearing in mind that our data comprise inputs utilized by firms to produce outputs (real values).

Support Vector Machines (Vapnik [1,16]) is a technique that stands out in ML in the world of supervised learning. SVM represents an algorithm constructed on the foundations of statistical learning theory and is in line with the Structural Risk Minimization (SRM) method. SRM is implemented to construct support vector machines, where the objective is to control the value of empirical risk and the value of the VC dimension, which is the regularization term that appears when the generalization error must be minimized rather than minimizing only the empirical error (Vapnik [1,16]). In particular, the definition of the notion of the VC dimension is as follows:

Definition 1 (VC dimension). *Let H be a set of binary-valued functions. A set of points X is shattered by H if for all binary vectors b indexed by X; there is a function $f_b \in H$ performing b on X. The VC dimension, $VC\ dim(H)$, of the set H is the size of the largest shattered set.*

With regard to the classification problem (for the regression problem, the generalization error is defined in the same way, because the regression problem can be turned in to a classification problem, as we will go on to explain) in SVM, minimizing the generalization error consists of minimizing the probability of incorrectly classifying any new data that emerges from the unknown distribution that was generated by the learning sample. This aim is possible if a bound of the generalization error is found, and the parameters on which it is dependent are controlled to reduce the bound. These bounds are understood as Probably Approximately Correct (PAC) bounds, which were first proposed by Valiant [17]. The standard PAC learning implements the idea of finding this classifier: it considers a fixed hypothesis (classifier) class together with a required accuracy and confidence, and takes into account the theory that characterizes when a function from this class can be learned from examples (data). In the case of regression, the exercise involves converting the regression problem (estimation function) into a classification problem because bounds in the generalization error are precisely based on the VC dimension or when a margin is considered on the fat-shattering dimension (effective VC dimension).

Next, we show the definition of the fat-shattering dimension. Notice that bold will be utilized for denoting vectors, and non-bold for scalars.

Definition 2 (fat-shattering dimension). *Let F be a set of real-valued functions. A set of points X is γ-shattered by F if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b indexed by X, there is a function $f_b \in F$ such that*

$$f_b(\boldsymbol{x}) \begin{cases} \geq r_x + \gamma & if\ b_x = 1 \\ \leq r_x - \gamma & otherwise \end{cases}.$$

The fat-shattering dimension of the set F, fat_F, is a function from the positive real numbers to the integers that maps a value γ to the largest γ-*shattered* set. The VC dimension corresponds to the largest shattered set, considering $\gamma = 0$, which is the concept first used by Vapnik to state a bound for the generalization error. This is the reason why the fat-shattering dimension is also known as the effective VC dimension.

To convert the regression problem into a classification problem, a threshold $\theta > 0$ that marks the limit needs to be set, such that a mistake will be considered to have been made if it is exceeded by the loss function when testing with new data in the model. The function that determines the distance between the real value of the output and the estimated value of said output through the model is called the loss function. Given a margin $\gamma > 0$, in the case of the training point, if the loss function exceeds the value $(\theta - \gamma)$, it will be considered as a mistake. Then, γ measures the discrepancy between the two losses: those measured on test data and those measured on training data. Under this re-interpretation of the regression problem, it is possible to use the dimension free bounds already constructed in the case of classification. In our case, we focus on the bound obtained by Shawe-Taylor and Cristianini [18], based on the fat-shattering dimension.

Theorem 1 (Shawe-Taylor and Cristianini [18]). *Let F be a sturdy class of real-valued functions with range $[-a, a]$ and fat-shattering dimension bounded by $fat_F(\gamma)$. Fix $\theta \in R$ with $\theta > 0$, and a scaling of the output range $\kappa \in R_+$. Consider a fixed but unknown probability distribution in the space $X \times R$. Then, with probability $1 - \rho$ over randomly drawn training sets S of size m for all γ with $\theta \geq \gamma > 0$, the probability that a training set filtered function $f \in F$ has an error larger than θ on a randomly chosen input is bounded by*

$$\varepsilon(m,d,\rho)=\frac{2}{m}\left(d\log_2\left(256m\left(\frac{c}{\gamma}\right)^2\right)\times\log_2\left(16em\left(\frac{c}{\gamma}\right)\right)+\log_2\left(\frac{16m^{1.5}a}{\rho\kappa}\right)\right) \quad (1)$$

where $c=\max\{a,D(S,f,\gamma)+\kappa\}$ *and*

$$d=\left[fat_F(\gamma^-/16)+\left(\frac{16(D(S,f,\gamma)+\kappa)}{\gamma}\right)^2\right], \quad (2)$$

provided $m\geq\frac{2}{\varepsilon}$.

In the statement of Theorem 1, $D(S,f,\gamma)=\sqrt{\sum_{(x,y)\in S}\xi((\boldsymbol{x},y),f,\gamma)^2}=\|\xi\|_2$ and $\xi((\boldsymbol{x},y),f,\gamma)=\max\{0,e(f)(\boldsymbol{x},y)-(\theta-\gamma)\}$, where $e(f)$ is the loss function that the analyst selects in order to measure how much f exceeds the error margin $(\theta-\gamma)$. In addition, the theorem introduces the concept of the fat-shattering dimension, $fat_F(\gamma)$, that is, the generalization of the VC dimension, which is sensitive to the size of the margin γ.

Theorem 1 is a general result, which in the case of each function class F, will be particularized: for each function class, the fat-shattering dimension is bounded in a different way, and consequently, the same happens with respect to the expected error proposed in (1). In the case of linear function classes, the fat-shattering dimension is bounded by Bartlett and Shawe-Taylor [19].

Theorem 2 (Bartlett and Shawe-Taylor [19]). *Suppose that* X *is a ball of radius* r *and center* $\boldsymbol{0}_m$ *in* R^m*, i.e.,* $X=\{\boldsymbol{x}\in R^m:\|\boldsymbol{x}\|\leq r\}$*, and consider the set*

$$F=\{\boldsymbol{x}\to\boldsymbol{w}\cdot\boldsymbol{x}:\|\boldsymbol{w}\|\leq 1,\boldsymbol{x}\in X\},$$

Then

$$fat_F(\gamma)\leq\left(\frac{r}{\gamma}\right)^2.$$

The most general version of this theorem, in which $\|\boldsymbol{w}\|$ is not restricted to be at most 1, bounds the fat-shattering dimension of linear classifiers as $fat_F(\gamma)\leq\left(\frac{\|\boldsymbol{w}\|r}{\gamma}\right)^2$.

The following two previously published lemmas are significant for our purposes throughout this paper (see Bartlett and Shawe-Taylor [19]).

Lemma 1. *For every input set* S *γ-shattered by* $F=\{\boldsymbol{x}\to\boldsymbol{w}\cdot\boldsymbol{x}:\boldsymbol{x}\in X\}$ *(the linear hypothesis class) and for every subset* $S_0\subseteq S$, $\|\sum S_0-\sum(S-S_0)\|\geq\frac{|S|\gamma}{\|\boldsymbol{w}\|}$ *holds.*

Lemma 2. *For all* $S\subseteq R_+^m$ *with* $\|\boldsymbol{x}\|\leq r$ *for* $\boldsymbol{x}\in S$*, certain* $S_0\subseteq S$ *satisfies that* $\|\sum S_0-\sum(S-S_0)\|\leq\sqrt{|S|}r$.

Then, $\frac{|S|\gamma}{\|\boldsymbol{w}\|}\leq\|\sum S_0-\sum(S-S_0)\|\leq\sqrt{|S|}r$. In particular, $\frac{|S|\gamma}{\|\boldsymbol{w}\|}\leq\sqrt{|S|}r$, and it is possible to conclude that all sets of inputs S *γ-shattered* by F are bounded. Therefore, the set *γ-shattered* by F with higher cardinality is also bounded, which is known as the fat-shattering dimension: $fat_F(\gamma)\leq\left(\frac{\|\boldsymbol{w}\|r}{\gamma}\right)^2$.

Now, if γ is fixed in such a way that $\theta\geq\gamma>0$, and disregarding the logarithmic factors in (1), the only term to reduce the expected error is (2). This process can be performed

by implementing the minimization of its bound, which in the case of linear functions, is as follows:

$$d = \left[fat_F(\gamma^-/16) + \left(\frac{\overbrace{16(D(S,f,\gamma)+\kappa)}^{D}}{\gamma} \right)^2 \right] \leq |\boldsymbol{w}|^2 + CD^2. \tag{3}$$

This expected error bound meets the SRM objective: the minimization process leads to more than minimizing the empirical risk, i.e., $D^2 = \sum_{(x,y)\in S} \xi((\boldsymbol{x},y),f,\gamma)^2$. Instead, it minimizes the capacity of the estimation function to provide a suitable prediction when a new observation (out of sample) is introduced and that is given by the appearance of the regularization term, that is $\|\boldsymbol{w}\|^2$, which bounds the fat-shattering dimension (PAC bound). The minimization of this bound corresponds to the objective of the regression problem associated with Support Vector Regression (SVR).

Support Vector Regression (SVR), as with any regression approach, attempts to construct a function that is capable of predicting the behavior of the response variable under the study. SVR sets out to predict the value of a continuous response variable $y \in R_+$ given a vector of covariables $\boldsymbol{x} \in R_+^m$. Hence, SVR establishes a function $\hat{f} : R_+^m \to R$ such that, given $\boldsymbol{x}$, $\hat{f}(\boldsymbol{x})$ yields the response variable prediction. Under the SVR principle, the linear predictor $\hat{f}$ can be defined as $\hat{f}(\boldsymbol{x}) = \boldsymbol{w}^* \cdot \boldsymbol{x} + b^*$, where $\boldsymbol{w}^* \in R^m$ and $b^* \in R$ are optimal solutions of the optimization model below:

$$\begin{array}{ll} \underset{\boldsymbol{w},b,\xi'_i,\xi_i}{Min} & \|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\left(\xi'^2_i + \xi_i^2\right) \\ & y_i - (\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \leq \varepsilon + \xi'_i, \quad i = 1,\ldots,n \\ & (\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1,\ldots,n \\ & \xi'_i, \xi_i \geq 0, \quad i = 1,\ldots,n \end{array} \tag{4}$$

In performing this methodology, the values of $C \in R_+$ and $\varepsilon \in R_+$ are obtained by a cross-validation process. The SVR yields an estimator $\hat{f}(\boldsymbol{x})$ of the response variable given $\boldsymbol{x}$ as well as lower and upper 'correcting' surfaces, defined as $\hat{f}(\boldsymbol{x}) - \varepsilon$ and $\hat{f}(\boldsymbol{x}) + \varepsilon$, where ε is a margin that enhances the estimator linked to SVR with robustness (see Figure 1). Additionally, observations below the surface $\hat{f}(\boldsymbol{x}) - \varepsilon$ reveal an associated (empirical) error of $\xi_i > 0$ (with $\xi'_i = 0$), while observations above the surface $\hat{f}(\boldsymbol{x}) + \varepsilon$ present an (empirical) error of $\xi'_i > 0$ (with $\xi_i = 0$). Observations between the surfaces $\hat{f}(\boldsymbol{x}) - \varepsilon$ and $\hat{f}(\boldsymbol{x}) + \varepsilon$ reveal an error of zero (with $\xi_i = \xi'_i = 0$). The objective function, however, represents the combination of regression and regularization involved in SVR, combining the empirical error term $\sum_{i=1}^{n}\left(\xi'^2_i + \xi_i^2\right)$ and the regularization term $\|\boldsymbol{w}\|^2$ through a weight C, thus balancing both components (Vazquez and Walter [20]). Moreover, although hyperplanes are linear in shape, it must be highlighted that SVR is able to generate estimation functions that are not necessarily linear in the original $(\boldsymbol{x}, y)$ space, and that can be achieved by using a transformation function ϕ, a conversion arising from the covariable space, $\phi : R_+^m \to Z$. Figure 1 shows the solution of the linear estimator achieved by an SVR model, as well as the graphical representation of the residuals (empirical error) for two points and the hyperplanes that define the margins (dashed lines).

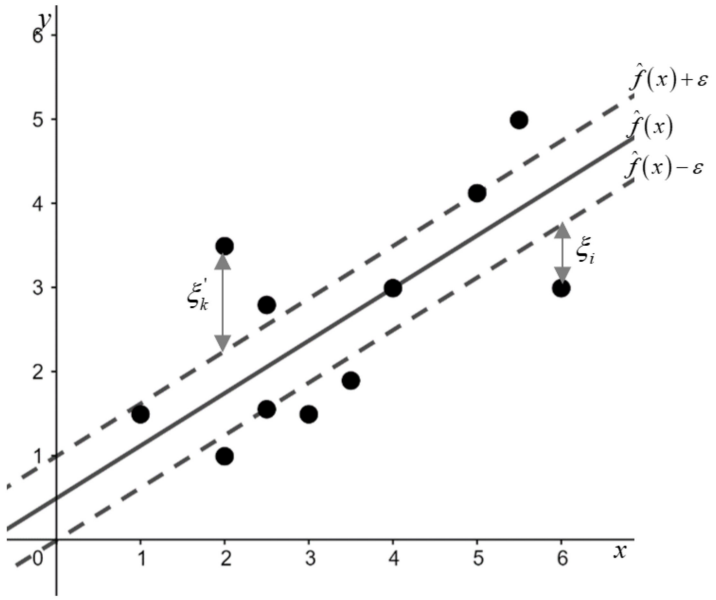

Figure 1. Support Vector Regression.

The next subsection explains how Data Envelopment Analysis (DEA) works.

2.2. Data Envelopment Analysis (DEA)

Let us consider the observation of n Decision Making Units (DMUs). DMU_i takes up $\boldsymbol{x}_i = \left(x_i^{(1)}, \ldots, x_i^{(m)}\right) \in R_+^m$ amounts of inputs to generate $\boldsymbol{y}_i = \left(y_i^{(1)}, \ldots, y_i^{(s)}\right) \in R_+^s$ amounts of outputs. The relative efficiency of each unit in the sample is evaluated by referring to the so-called production possibility set or technology, which is essentially the set of producible bundles of $(\boldsymbol{x}, \boldsymbol{y})$. It is generally defined as:

$$T = \left\{(\boldsymbol{x}, \boldsymbol{y}) \in R_+^{m+s} : \boldsymbol{x} \text{ can produce } \boldsymbol{y}\right\} \tag{5}$$

Under Data Envelopment Analysis (DEA) (Charnes et al. [5] and Banker et al. [6] and more recently, Villa et al. [21], Sahoo et al. [22], and Amirteimoori [23]), T is usually assumed to satisfy free disposability with regard to inputs and outputs; that is, if $(\boldsymbol{x}, \boldsymbol{y}) \in T$, then $(\boldsymbol{x}', \boldsymbol{y}') \in T$ with $\boldsymbol{x}' \geq \boldsymbol{x}$ and $\boldsymbol{y}' \leq \boldsymbol{y}$. Convexity of T is also generally assumed (see, e.g., Färe and Primont [24]).

Insomuch as the measurement of technical efficiency is concerned, a certain subset of T is of interest. We allude to the weakly efficient set of T, defined as $\partial^W(T) := \{(\boldsymbol{x}, \boldsymbol{y}) \in T : \hat{\boldsymbol{x}} < \boldsymbol{x}, \hat{\boldsymbol{y}} > \boldsymbol{y} \Rightarrow (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \notin T\}$ (Let $\boldsymbol{z} = \left(z^{(1)}, \ldots, z^{(q)}\right)$ and $\boldsymbol{t} = \left(t^{(1)}, \ldots, t^{(q)}\right)$. Then, $\boldsymbol{z} < \boldsymbol{t}$ means $z^{(j)} < t^{(j)}$ for all $j = 1, \ldots, q$.). Some authors (see, for example, Briec and Lesourd [25]) define technical efficiency as the distance from a point in T to the weakly efficient set.

When $s = 1$, this context is confined to the central concept of production function f. Accordingly, m input variables are used to yield a univariate output, and hence, we can define the technology as:

$$T = \left\{(\boldsymbol{x}, y) \in R_+^{m+1} : y \leq f(\boldsymbol{x})\right\}.$$

According to the selected distance for measuring technical inefficiency, different DEA models emerge (Cooper et al. [26]). The directional distance function (DDF) is a relevant example of them. For m inputs and one output, resorting to the directional vector

$\boldsymbol{g} = (\boldsymbol{g}^-, g^+)$, where $\boldsymbol{g}^- = \mathbf{1}_m$ and $g^+ = 1$, the DDF problem has the following structure when the efficiency level of DMU_i is assessed, $i = 1, \ldots, n$:

$$\begin{array}{ll} \underset{\beta_i, \lambda_1, \ldots, \lambda_n}{Max} & \beta_i \\ s.t. & \\ & \sum_{k=1}^{n} \lambda_k x_k^{(j)} \leq x_i^{(j)} - \beta_i, \quad \forall j = 1, \ldots, m \\ & \sum_{k=1}^{n} \lambda_k y_k \geq y_i + \beta_i, \\ & \sum_{k=1}^{n} \lambda_k = 1, \\ & \lambda_k \geq 0 \qquad\qquad \forall k = 1, \ldots, n \end{array} \tag{6}$$

Given that (6) is a linear program, we can equivalently solve its corresponding dual formulation:

$$\begin{array}{ll} \underset{\boldsymbol{c}_i, p_i, \alpha_i}{Min} & -p_i y_i + \boldsymbol{c}_i \boldsymbol{x}_i + \alpha_i \\ s.t. & \\ & p_i y_k - \boldsymbol{c}_i \boldsymbol{x}_k - \alpha_i \leq 0, \quad \forall k = 1, \ldots, n \\ & \|(\boldsymbol{c}_i, p_i)\|_1 = 1, \\ & p_i \geq 0, \\ & c_i^{(j)} \geq 0, \qquad\qquad \forall j = 1, \ldots, m \end{array} \tag{7}$$

DEA models must be solved for each DMU_i, $i = 1, \ldots, n$, in the sample.

Figure 2 shows an example of the DDF model with a distance vector $\boldsymbol{g} = (\boldsymbol{g}^-, g^+) = (\mathbf{1}_m, 1)$. Note that DEA generates a piece-wise linear technology (the region below the line), satisfying free disposability in inputs and outputs and convexity. Note also that the DEA estimate envelops all the observations from above. In this case, with $\boldsymbol{g} = (\mathbf{1}_m, 1)$, the DDF coincides with a particular distance between data and $\partial^W(T)$: the l_∞-distance (Briec [27] and Briec and Lesourd [25]).

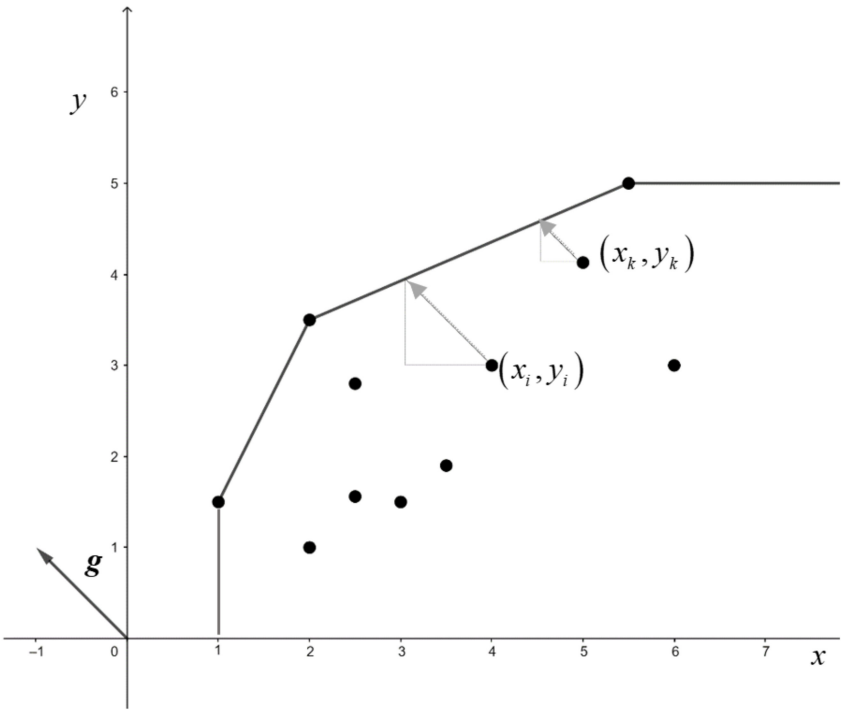

Figure 2. Illustration of the Directional Distance Function in Data Envelopment Analysis.

In this paper, our purpose is to construct a method that generates piece-wise linear frontiers as in Figure 2, by implementing the minimization of the generalization error of the model.

3. New PAC Learning with Piece-Wise Linear Hypothesis

This section revolves around two stages in the search for the generalization error bound: the first stage is based on the construction of the class of piecewise linear hypotheses whose elements are hyperplanes that are located as close as possible to the data sample through l_∞-*distance*, and the second stage is based on the construction of the bound of the fat-shattering dimension of the class of hypothesis constructed in the first stage. The minimization of the bound of the expected error using the bound of the fat-shattering dimension calculated gives rise to the Data Envelopment Analysis-based Machines (DEAM) model as a method for estimating piecewise linear production functions, which minimizes the generalization error as well as the empirical error.

To obtain this bound of the class of functions of our interest, we must derive the fat-shattering dimension bound for the hypothesis class with the piece-wise structure we desire. Then, minimizing the generalization error will be implemented through the minimization of the fat-shattering dimension bound. For this task, a previous step must be taken: a class of piece-wise linear hypothesis must be defined. A piece-wise linear hypothesis target is defined by a combination of n hyperplanes $\left\{H_p\right\}_{p=1,\ldots,n}$ that are selected to evaluate the data depending on their input values. The hyperplanes will be defined for each input value $\boldsymbol{x} \in R_+^m$ as follows:

$$H_{p_x} = \left\{(\boldsymbol{x}, y) \in R^{m+1} : \boldsymbol{w}_{p_x}\boldsymbol{x} + \beta_{p_x} - \delta_{p_x} y = 0\right\}$$

Then, if we suppose $\delta_{p_x} > 0, \forall p_x \in \{1, \ldots, n\}$, each output value estimation through the set of n hyperplanes $\left\{H_p\right\}_{p=1,\ldots,n}$ can be written as a function of the input value vector $\boldsymbol{x}$:

$$h(\boldsymbol{x}) = \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}}, \quad p_x \in \{1, \ldots, n\} \ ,$$

with $\boldsymbol{w}_{p_x} \in R_+^m$ and $\beta_{p_x} \in R$. The value of $p_x \in \{1, \ldots, n\}$, in our case, is chosen by considering two desired conditions that are inherited from production theory:

$$h(\boldsymbol{x}) = \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \geq 0, \tag{8}$$

and

$$\frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \leq \frac{\boldsymbol{w}_p}{\delta_p}\boldsymbol{x} + \frac{\beta_p}{\delta_p}, \quad \forall p \in \{1, \ldots, n\} \ . \tag{9}$$

Condition (8) ensures that the estimation of the output value associated with an input $\boldsymbol{x} \in R_+^m$ will be always non-negative. Additionally, condition (9) guarantees that the estimation $h(\boldsymbol{x})$ through the hyperplane H_{p_x} is less or equal than the estimation through any other hyperplane H_p. Condition (9) is the one that imposes concavity on the model. This type of condition was the key for stating concavity in the general multiple-regressor modeling in microeconomics (Afriat [28]; Kuosmanen et al. [13]). In particular, if the production function is concave, then the technology defined from this production function is convex.

The function class of piece-wise linear hypothesis can be constructed as follows:

$$\begin{aligned} F = \quad & \left\{x \mapsto \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} : \|\boldsymbol{x}\| \leq R, \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \geq 0,\right. \\ & \left.\frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \leq \frac{\boldsymbol{w}_p}{\delta_p}\boldsymbol{x} + \frac{\beta_p}{\delta_p}, \forall p_x, p \in \{1, \ldots, n\}\right\}. \end{aligned} \tag{10}$$

Now, we can proceed with the second step: to establish a bound for the fat-shattering dimension of this function class to control the generalization error. Before proving the main

theorem of this section, we need to state a necessary technical lemma. In the results, $r \in R_+$ is the radius of the ball centered in $\mathbf{0}_m$ that bounds the input data in the data sample.

Lemma 3. *If an input learning sample, $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ is γ-shattered through F defined in (10), then every subset $S_0 \subseteq S$ satisfies*

$$\left\| \sum S_0 - \sum (S - S_0) \right\| \geq |S| \left(\frac{\underset{p \in \{1,\ldots,n\}}{Min} \left\{ \gamma - \frac{\beta_p}{\delta_p} \right\}}{\underset{p \in \{1,\ldots,n\}}{Max} \left\| \frac{\boldsymbol{w}_p}{\delta_p} \right\|} - 2r \right), \tag{11}$$

denoting as $\sum S_0$ and $\sum (S - S_0)$ the sum of the elements in S_0 and $S - S_0$, respectively, and as $|S|$ the cardinal of the set S.

Proof. See Appendix A. □

Next, we prove the main theorem of this section. In particular, we state the bound for the fat-shattering dimension for piece-wise linear hypothesis classes.

Theorem 3. *Let X be the ball of radius R and center $\mathbf{0}_m$ in R^m, i.e., $X = \{\boldsymbol{x} \in R^m : \|\boldsymbol{x}\| \leq r\}$, and let the hypothesis class be as follows*

$$\begin{aligned} F = \Bigg\{ x \mapsto \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}} \boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} : \|x\| \leq R, \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}} \boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \geq 0, \\ \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}} \boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \leq \frac{\boldsymbol{w}_p}{\delta_p} \boldsymbol{x} + \frac{\beta_p}{\delta_p}, \forall p_x, p \in \{1, \ldots, n\} \Bigg\}, \end{aligned} \tag{12}$$

then

$$fat_F(\gamma) \leq \left(\frac{r}{\frac{\underset{p \in \{1,\ldots,n\}}{Min} \left\{ \gamma - \frac{\beta_p}{\delta_p} \right\}}{\underset{p \in \{1,\ldots,n\}}{Max} \left\| \frac{\boldsymbol{w}_p}{\delta_p} \right\|} - 2r} \right)^2. \tag{13}$$

Proof. See Appendix A. □

The next section involves the task of achieving a model that minimizes the established generalization error through the l_∞-distance.

4. Data Envelopment Analysis-Based Machines (DEAM)

Data Envelopment Analysis-based Machines (DEAM) can be defined from the idea of minimizing the expected error proposed in (1). If we do not consider the logarithmic factors, we can directly focus on minimizing d in this expression, for which a bound on the generalization error has been found in the case of the piece-wise linear hypothesis class F defined in (12):

$$d = fat_F(\gamma^-/16) + \left(\frac{16(D(S, f, \gamma) + \kappa)}{\gamma} \right)^2 \leq \underbrace{\left(\frac{r}{\frac{\underset{p \in \{1,\ldots,n\}}{Min} \left\{ \frac{\gamma}{16} - \frac{\beta_p}{\delta_p} \right\}}{\underset{p \in \{1,\ldots,n\}}{Max} \left\| \frac{\boldsymbol{w}_p}{\delta_p} \right\|} - 2r} \right)^2}_{A} + \underbrace{\left(\frac{16(D(S, f, \gamma) + \kappa)}{\gamma} \right)^2}_{B} \tag{14}$$

Because of the complexity of implementing an optimization model in which the objective function has the aim of minimizing the above bound, we will break up the minimization of the whole bound into different objectives, which will be collected in an aggregation function that will conform the objective function of the final optimization program associated with DEAM, which will be shown later in this section.

Once the number of different hyperplanes in each hypothesis is set as the number of elements in the learning sample $|S| = n$, minimizing the bound of the fat-shattering dimension requires minimizing part A in (14). This is equivalent to maximizing $\frac{\underset{p\in\{1,\dots,n\}}{Min}\left\{\frac{\gamma}{16}-\frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\dots,n\}}{Max}\left\|\frac{\boldsymbol{w}_p}{\delta_p}\right\|}$. Regarding this last expression, we must maximize the numerator and minimize the denominator, as follows:

(i) The vector of coefficients (slopes) corresponding to the hyperplane H_p is $(\boldsymbol{w}_p, \delta_p)$. We can consider, without loss of generality, that $\|(\boldsymbol{w}_p, \delta_p)\|_1 = 1$. Then, minimizing $\underset{p\in\{1,\dots,n\}}{Max}\left\|\frac{\boldsymbol{w}_p}{\delta_p}\right\|$, is equivalent to minimizing $\underset{p\in\{1,\dots,n\}}{Max}\left(\frac{1}{\frac{1}{\|\boldsymbol{w}_p\|}-1}\right)$ since $\delta_p \geq 0$, $\forall p \in \{1,\dots,n\}$. Focusing on that last equivalence, this objective can be directly translated into minimizing $\underset{p\in\{1,\dots,n\}}{Max}\|\boldsymbol{w}_p\|$.

(ii) Maximizing $\underset{p\in\{1,\dots,n\}}{Min}\left\{\frac{\gamma}{16}-\frac{\beta_p}{\delta_p}\right\}$ with a fixed value of the margin γ is equivalent to minimizing $\underset{p\in\{1,\dots,n\}}{Max}\left\{\frac{\beta_p}{\delta_p}\right\}$. Because of $\|(\boldsymbol{w}_p, \delta_p)\|_1 = 1$, by minimizing $\underset{p\in\{1,\dots,n\}}{Max}\|\boldsymbol{w}_p\|$ in (i), at the same time, the maximization of the elements $\{\delta_p\}_{p\in\{1,\dots,n\}}$ is achieved. In this way, it is only necessary to minimize $\underset{p\in\{1,\dots,n\}}{Max}\{\beta_p\}$ to maximize $\underset{p\in\{1,\dots,n\}}{Max}\left\{\frac{\beta_p}{\delta_p}\right\}$.

Finally, a way of implementing (i) and (ii) is minimizing $u + v$, where $\|\boldsymbol{w}_p\| \leq u$, $\beta_p \leq v, \forall p \in \{1,\dots,n\}$. Accordingly, minimizing the bound of the fat-shattering dimension, $A + B$, leads to minimizing $(u + v) + CD^2$, where $D^2 = D(S, f, \gamma)^2 = \|\xi\|_2^2$ and C is a parameter to be tuned by, for example, a cross-validation process. As a loss function we use the following: $\xi((\boldsymbol{x}, y), f, \gamma) = \max\left\{0, D_{\|\cdot\|_\infty}((\boldsymbol{x}, y), f) - (\theta - \gamma)\right\}$. Finally, the objective function has the following structure:

$$z(u, v, \xi_1, \dots, \xi_n) = u + v + C\|\xi\|_2^2. \tag{15}$$

Accordingly, we introduce the optimization model that defines DEAM:

$$\begin{aligned}
\underset{\substack{w,\beta,\delta,\xi,\xi',u,v\\ s.t.}}{Min}\quad & u + v + C\left(\sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\xi_i'^2\right) & & \\
& \|\boldsymbol{w}_i\|_1 \leq u & \forall i = 1,\dots,n & (16.1)\\
& \beta_i \leq v & \forall i = 1,\dots,n & (16.2)\\
& \delta_p y_i \leq \boldsymbol{w}_p \boldsymbol{x}_i + \beta_p & \forall i, p = 1,\dots,n & (16.3)\\
& \boldsymbol{w}_i, \delta_i \geq 0 & \forall i = 1,\dots,n & (16.4)\\
& \boldsymbol{w}_i \boldsymbol{x}_i + \beta_i - \delta_i y_i \leq \varepsilon + \xi_i & \forall i = 1,\dots,n & (16.5)\\
& \delta_i y_i - \boldsymbol{w}_i \boldsymbol{x}_i - \beta_i \leq \varepsilon + \xi_i' & \forall i = 1,\dots,n & (16.6)\\
& \xi_i, \xi_i' \geq 0 & \forall i = 1,\dots,n & (16.7)\\
& \|(\boldsymbol{w}_i, \delta_i)\|_1 = 1 & \forall i = 1,\dots,n & (16.8)\\
& \boldsymbol{w}_i \boldsymbol{x}_i + \beta_i - \delta_i y_i \leq \boldsymbol{w}_p \boldsymbol{x}_i + \beta_p - \delta_p y_i & \forall i, p = 1,\dots,n & (16.9)
\end{aligned} \tag{16}$$

Model (16) determines a maximum of n different hyperplanes. The intersection of the half-spaces defined from these hyperplanes gives rise to the estimator of the underlying (convex) production technology. The number of hyperplanes to be considered in the implementation of the DEAM model can be seen as a key parameter of our approach since

the results could be different depending on it. However, we suggest using n hyperplanes, which coincide with the number of DMUs. This is due to the experimental evidence found in the simulation study carried out in Section 5. We analyzed 2000 databases, and in all these cases, the number of hyperplanes at optimum were less than the number of DMUs in the corresponding data sample. This situation can be identified because some hyperplanes are repeated at the optimal solution of each problem.

Let us now explain each constraint of model (16) in detail. Constraints (16.1) and (16.2) come from $\|\boldsymbol{w}_p\| \leq u$, $\beta_p \leq v, \forall p = 1, \ldots, n$, respectively. The norm l_1 is used to be consistent with constraint (16.8). Additionally, this type of norm is associated with the definition of linear constraints, which are easier to be solved from a computational point of view. Constraint (16.3) is equivalent to $y_i \leq \frac{\boldsymbol{w}_p}{\delta_p}\boldsymbol{x}_i + \frac{\beta_p}{\delta_p}$, $i, p = 1, \ldots, n$, i.e., it ensures that the hyperplanes envelop the data sample from above. Condition (16.4) forces that the n hyperplanes are monotonic non-decreasing and will be responsible for the satisfaction of the property of free disposability, as we will show later in the text (see Proposition 2 below). Constraints (16.5), (16.6), (16.7), and (16.8) allow for characterizing $\xi((\boldsymbol{x}, y), f, \gamma)$ as $\max\left\{0, D_{\|\cdot\|_\infty}((\boldsymbol{x}, y), f) - (\theta - \gamma)\right\}$. The parameter ε $(= \theta - \gamma \geq 0)$ will be chosen by cross validation. Let us now interpret specifically the value at optimum of the decision variable ξ_i. Let us pay attention to constraint (16.5). If $\boldsymbol{w}_i\boldsymbol{x}_i + \beta_i - \delta_i y_i - \varepsilon \geq 0$, then $\xi_i = \boldsymbol{w}_i\boldsymbol{x}_i + \beta_i - \delta_i y_i - \varepsilon$ since $\sum_{i=1}^{n} \xi_i^2$ is minimized in the objective function. In this way, considering (16.8), (16.3) and $\varepsilon \geq 0$, ξ_i can be interpreted as the l_∞-distance from the observation $(\boldsymbol{x}_i, y_i)$ to the hyperplane $H_{i\varepsilon}$:

$$\xi_i = \frac{\left|\boldsymbol{w}_i\boldsymbol{x}_i + \beta_i - \delta_i y_i - \varepsilon\right|}{\left\|(\boldsymbol{w}_i, \delta_i)\right\|_1} = D_{l_\infty}((\boldsymbol{x}_i, y_i), H_{i\varepsilon}),$$

where $H_{i\varepsilon} = \left\{(\boldsymbol{x}, y) \in R^{m+1} : \boldsymbol{w}_i\boldsymbol{x} + \beta_i - \delta_i y - \varepsilon = 0\right\}$ (Mangasarian [29]). If $\boldsymbol{w}_i\boldsymbol{x}_i + \beta_i - \delta_i y_i - \varepsilon < 0$, then $\xi_i = 0$ by (16.7) and the minimization of $\sum_{i=1}^{n} \xi_i^2$. Additionally, regarding the value of ξ_i', $i = 1, \ldots, n$, by constraints (16.3), (16.6), $\varepsilon \geq 0$ and the minimization of $\sum_{i=1}^{n} \xi_i'^2$, we obtain $\xi_i' = 0$ for all $i = 1, \ldots, n$ at optimum. This point has computational implications on the model since constraint (16.6) can be removed from it because (16.3) holds. Finally, constraint (16.9) guarantees that, for each $(\boldsymbol{x}_i, y_i)$ in the data sample, the hyperplane of the piece-wise linear production function associated with that point is the closest one to $(\boldsymbol{x}_i, y_i)$. Note that constraint (16.9), by (16.3) and (16.8), is equivalent to writing $D_{l_\infty}((\boldsymbol{x}_i, y_i), H_i) \leq D_{l_\infty}\left((\boldsymbol{x}_i, y_i), H_p\right)$ $\forall i, p = 1, \ldots, n$ (see Mangasarian [29]).

Figure 3 shows the shape of the function that will be generated by the model as an estimate of the underlying production function. Note that the estimate satisfies monotonicity and concavity, as happens with the DEA estimator. However, the DEAM estimator does not satisfy minimal extrapolation. Additionally, it implements a certain idea of robustness because of the margin notion inherited from SVR. Additionally, Figure 3 shows the possible interpretation of ξ_i as $D_{l_\infty}((\boldsymbol{x}_i, y_i), H_{i\varepsilon})$. In particular, ξ_i is the 'radius' of the squared ball in the figure.

As the technology generated by DEA, DEAM provides a piece-wise linear technology that can be defined as $T_{DEAM} := \left\{(\boldsymbol{x}, y) \in R_+^{m+1} : \boldsymbol{w}_p^*\boldsymbol{x} + \beta_p^* - \delta_p^* y \geq 0, \forall p \in \{1, \ldots, n\}\right\}$, given an optimal solution $\left(\left\{\boldsymbol{w}_p^*, \beta_p^*, \delta_p^*, \xi_p^*, \xi_p^{*\prime}\right\}_{p=1,\ldots,n}, u^*, v^*\right)$ of model (16).

The next propositions state that the derived technology from model (16) satisfies convexity and free disposability.

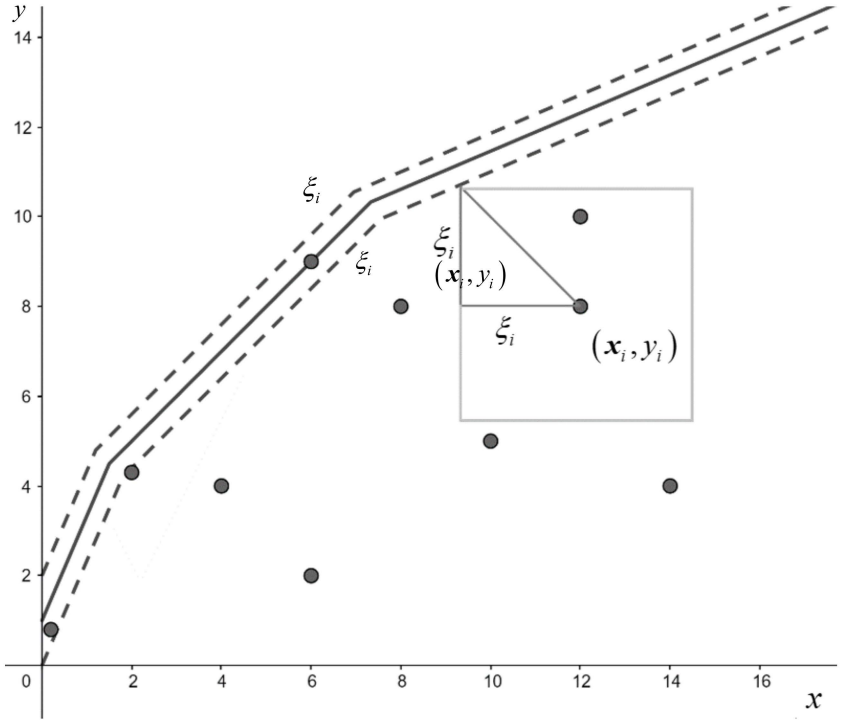

Figure 3. Illustration of the DEAM estimation of a production function.

Proposition 1. *T_{DEAM} is a convex set.*

Proof. The intersection of half-spaces is a convex set. □

Proposition 2. *T_{DEAM} satisfies free disposability in inputs and outputs.*

Proof. The result holds because $\boldsymbol{w}_p, \delta_p \geq 0$ for all $p \in \{1, \ldots, n\}$ (see Kuosmanen and Johnson [13]). □

Additionally, by constraint (16.3), we have that $\boldsymbol{w}_p^* \boldsymbol{x}_i + \beta_p^* - \delta_p^* y_i \geq 0, \forall i, p = 1, \ldots, n$. Therefore, for any observation $(\boldsymbol{x}_{i'}, y_{i'})$, we have that $\boldsymbol{w}_p^* \boldsymbol{x}_{i'} + \beta_p^* - \delta_p^* y_{i'} \geq 0, \forall p = 1, \ldots, n$, which implies that $(\boldsymbol{x}_{i'}, y_{i'}) \in T_{DEAM}$ since $T_{DEAM} = \left\{ (\boldsymbol{x}, y) \in R_+^{m+1} : \boldsymbol{w}_p^* \boldsymbol{x} + \beta_p^* - \delta_p^* y \geq 0, \right.$ $\left. \forall p \in \{1, \ldots, n\} \right\}$. In this way, we can establish the following corollary.

Corollary 1. *The production possibility set generated by DEA is a subset of the production possibility set generated by DEAM.*

Proof. The result holds because the production possibility set generated by DEA and the production possibility set yielded by DEAM satisfy convexity, free disposability, and contain all observations, but only the technology related to DEA meets minimal extrapolation. □

In this way, we have that DEAM does not satisfy the minimal extrapolation principle, but its associated estimation of the technology always contains the observations.

As for the measurement of technical inefficiency of the observations, due to the nature of the technique used and based on the original ideas derived from Support Vector Regression, any $(\boldsymbol{x}_i, y_i)$ located within the margin will be identified as technically efficient (with $\xi_i^* = 0$). Otherwise, i.e., if $(\boldsymbol{x}_i, y_i)$ is located below the margin (see Figure 3), we have that ξ_i^* is the l_∞-distance from the observation to the (efficient) frontier of a 'robust' technology. This robust technology is defined by the translation of the original technol-

ogy T_{DEAM} downward following the value of the margin ε. If we define this translated technology as $T^{\varepsilon}_{DEAM} = \left\{(\boldsymbol{x}, y) \in R^{m+1}_{+} : \boldsymbol{w}^{*}_{p}\boldsymbol{x} + \beta^{*}_{p} - \delta^{*}_{p}y - \varepsilon \geq 0, \forall p \in \{1, \ldots, n\}\right\}$, then $\xi^{*}_{i} = D_{l_\infty}\left((\boldsymbol{x}_i, y_i), \partial^{W}\left(T^{\varepsilon}_{DEAM}\right)\right)$ (this result can be derived from Aparicio and Pastor [30]).

Now, we show the relationship between the Directional Distance Function (DDF) in DEA, model, and the DEAM model (16): The DDF model always yields a feasible solution of the model associated with Data Envelopment Analysis-based Machines.

Theorem 4. *Let* $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}\right)\right\}_{i=1,\ldots,n}$ *be a set of optimal solutions of model (7) for each* DMU_i, $i = 1, \ldots, n$. *Then,* $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}, \vartheta^{*}_{i}, \vartheta'^{*}_{i}\right)_{i=1,\ldots,n}, a^{*}, b^{*}\right\}$, *with* $\vartheta^{*}_{i} = -p^{*}_{i}y_i + \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i + \alpha^{*}_{i}$, $\vartheta'^{*}_{i} = p^{*}_{i}y_i - \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i - \alpha^{*}_{i} = 0, \forall i = 1, \ldots, n$, $a^{*} = \max\limits_{i=1,\ldots,n}\left\|\boldsymbol{c}^{*}_{i}\right\|$, $b^{*} = \max\limits_{i=1,\ldots,n}\left\{\alpha^{*}_{i}\right\}$ *is a feasible solution of model (16).*

Proof. Let $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}\right)\right\}_{i=1,\ldots,n}$ be a set of optimal solutions of model (7) for each DMU_i, $i = 1, \ldots, n$. By the characterization of a^{*} and b^{*} as $a^{*} = \max\limits_{i=1,\ldots,n}\left\|c^{*}_{i}\right\|$ and $b^{*} = \max\limits_{i=1,\ldots,n}\left\{\alpha^{*}_{i}\right\}$, the following inequalities are true:

$$\left\|\boldsymbol{c}^{*}_{i}\right\| \leq a^{*} \tag{17}$$

$$\alpha^{*}_{i} \leq b^{*} \tag{18}$$

Then, $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}, \vartheta^{*}_{i}, \vartheta'^{*}_{i}\right)_{i=1,\ldots,n}, a^{*}, b^{*}\right\}$ satisfies (16.1) and (16.2) in the DEAM model. Because of the fact that $\left\{\left(c^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}\right)\right\}_{i=1,\ldots,n}$ is a set of optimal solutions of model (7) for each DMU_i, $i = 1, \ldots, n$, the constraints of this model are satisfied for this solution:

$$\begin{aligned} &p^{*}_{i}y_k \leq \boldsymbol{c}^{*}_{i}\boldsymbol{x}_k + \alpha^{*}_{i} && \forall k = 1, \ldots, n; \forall i = 1, \ldots, n && (19.1)\\ &\left\|\left(\boldsymbol{c}^{*}_{i}, p^{*}_{i}\right)\right\|_1 = 1 && \forall i = 1, \ldots, n && (19.2)\\ &\boldsymbol{c}^{*}_{i}, p^{*}_{i} \geq 0 && \forall i = 1, \ldots, n && (19.3) \end{aligned} \tag{19}$$

Then, $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}, \vartheta^{*}_{i}, \vartheta'^{*}_{i}\right)_{i=1,\ldots,n}, a^{*}, b^{*}\right\}$ trivially satisfies (16.3), (16.4) and (16.8) in the DEAM model. Because of the definition of the variables ϑ^{*}_{i} and ϑ'^{*}_{i} as $\vartheta^{*}_{i} = -p^{*}_{i}y_i + \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i + \alpha^{*}_{i}$, $\vartheta'^{*}_{i} = p^{*}_{i}y_i - \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i - \alpha^{*}_{i} = 0, \forall i = 1, \ldots, n$, we have that:

$$-p_iy_i + \boldsymbol{c}_i\boldsymbol{x}_i + \alpha_i \leq \vartheta^{*}_{i} + \varepsilon \tag{20}$$

and

$$p^{*}_{i}y_i - \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i - \alpha^{*}_{i} \leq \vartheta'^{*}_{i} + \varepsilon, \tag{21}$$

$\forall i = 1, \ldots, n$, and $\forall \varepsilon \geq 0$. Then, $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}, \vartheta^{*}_{i}, \vartheta'^{*}_{i}\right)_{i=1,\ldots,n}, a^{*}, b^{*}\right\}$ satisfies (16.5) and (16.6) in the DEAM model. Additionally, we have

$$0 \leq -p^{*}_{i}y_i + \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i + \alpha^{*}_{i} = \vartheta^{*}_{i} \quad \forall i = 1, \ldots, n \tag{22}$$

and,

$$0 \leq p^{*}_{i}y_i - \boldsymbol{c}^{*}_{i}\boldsymbol{x}_i - \alpha^{*}_{i} = \vartheta'^{*}_{i} \quad \forall i = 1, \ldots, n\ . \tag{23}$$

Constraint (22) is satisfied by (19.1), and (23) is trivially satisfied. Then, $\left\{\left(\boldsymbol{c}^{*}_{i}, \alpha^{*}_{i}, p^{*}_{i}, \vartheta^{*}_{i}, \vartheta'^{*}_{i}\right)_{i=1,\ldots,n}, a^{*}, b^{*}\right\}$ satisfies (16.7) in the DEAM model. Finally, the objective in (7) is to minimize $\vartheta_i = -p_iy_i + c_ix_i + \alpha_i, \forall i = 1, \ldots, n$,

that implies

$$\vartheta_i^* = -p_i^* y_i + \boldsymbol{c}_i^* \boldsymbol{x}_i + \alpha_i^* \leq -p_k y_i + \boldsymbol{c}_k \boldsymbol{x}_i + \alpha_k \quad \forall k = 1, \ldots, n \ . \tag{24}$$

Then, $\left\{ \left(\boldsymbol{c}_i^*, \alpha_i^*, p_i^*, \vartheta_i^*, \vartheta_i'^* \right)_{i=1,\ldots,n}, a^*, b^* \right\}$ satisfies (16.9) in the DEAM model. Consequently, $\left\{ \left(\boldsymbol{c}_i^*, \alpha_i^*, p_i^*, \vartheta_i^*, \vartheta_i'^* \right)_{i=1,\ldots,n}, a^*, b^* \right\}$ is a feasible solution of (16). □

However, it can be shown that the DDF model (7) does not always yield an optimal solution of model (16).

5. Computational Experience

This section compares the performance of DEA and DEAM for estimating production functions. For this task, we designed five typical production scenarios in Table 1.

Table 1. Simulated scenarios.

Scenario	Inputs	Production Function
I	x_1	$y = x_1^{0.5}$
II	x_1, x_2	$y = x_1^{0.35} \cdot x_2^{0.15}$
III	x_1, x_2, x_3	$y = x_1^{0.30} \cdot x_2^{0.15} \cdot x_3^{0.05}$
IV	x_1, x_2, x_3, x_4	$y = x_1^{0.25} \cdot x_2^{0.15} \cdot x_3^{0.05} \cdot x_4^{0.05}$
V	x_1, x_2, x_3, x_4, x_5	$y = x_1^{0.25} \cdot x_2^{0.10} \cdot x_3^{0.05} \cdot x_4^{0.05} \cdot x_5^{0.05}$

The simulations implement Cobb–Douglas production functions, which are frequently used in econometrics for establishing the relation between the maximum amount of outputs that can be produced from a set of inputs. Thereby, scenario I implements a mono-input mono-output case, while the other scenarios represent multi-input mono-output cases. For each scenario, we ran 100 trials $(t = 1, \ldots, 100)$ with sample sizes: $n \in \{25, 50, 75, 100\}$. The inputs were calculated randomly from $Uni[1, 10]$. For simulating inefficiencies, we selected a random distribution $\exp(1/3)$ for u. Mean squared error (MSE) and bias were the two measures employed to assess the performance of each method.

The DEAM model (16), as other machine learning techniques, needs to find the best model through a cross-validation process. For this task and exclusively for the DEAM model, we implemented a five-fold cross validation using a certain grid of hyperparameters. This grid was arbitrarily set as: $C \in \{1, 10, 50, 100, 10^6\}$ and $\varepsilon \in \{0, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 1\}$. Note that DEA does not need to apply a cross-validation process. Instead, DEA uses the whole dataset to evaluate efficiency scores.

Table 2 sums up the results obtained for each scenario when DEA (without cross validation) and DEAM (with cross validation) are applied. The first two columns present the type of scenario and the sample size. The following columns show the mean and standard deviation (in brackets) of MSE obtained by DEA and DEAM. Fraction of trial reports the proportion of trials in which DEAM either improves upon or equals the MSE given by the DEA method, while the next column illustrates the percentage of improvement of DEAM with respect to DEA. The four subsequent columns are similar to the previous ones, but with regard to bias.

Regarding the results, the DEAM method performed better than DEA, with improvements ranging from 5% to 45% on average in MSE and 2% to 28% in bias. This fact increased when the number of inputs were higher. In addition, the results illustrate how the model worked better when the number of DMUs was around 50–75. Scenario I, i.e., the single input single output framework, shows small differences between the two methods. Nevertheless, in the trials, DEAM outperformed DEA in more than 95% of the cases. In contrast, the best analyzed situation was scenario V (one output and five inputs) with $n = 25$, showing a 45% reduction in MSE and 28% in bias, on average. This last result could be interpreted in favor of the DEAM approach as an indication that DEAM also seemed to outperform DEA with respect to the curse of dimensionality (Charles et al. [31]).

Table 2. Performance of DEA and DEAM.

		MSE				BIAS			
				Fraction of Trials	Improvement (%)			Fraction of Trials	Improvement (%)
Scenario	**Number of Obs.**	**DEA**	**DEAM**	**DEAM<= DEA**	**DEAM vs. DEA**	**DEA**	**DEAM**	**DEAM<= DEA**	**DEAM vs. DEA**
I	**25**	0.027(0.020)	0.024(0.019)	1.000	11.609%	0.125(0.046)	0.119(0.046)	1.000	4.873%
I	**50**	0.011(0.007)	0.010(0.007)	0.990	8.005%	0.076(0.026)	0.075(0.026)	0.990	2.822%
I	**75**	0.007(0.005)	0.007(0.005)	0.990	7.622%	0.060(0.019)	0.059(0.019)	0.980	2.194%
I	**100**	0.005(0.004)	0.005(0.004)	0.990	5.231%	0.051(0.019)	0.050(0.019)	0.950	1.936%
II	**25**	0.151(0.084)	0.108(0.067)	1.000	27.109%	0.276(0.071)	0.240(0.072)	1.000	13.460%
II	**50**	0.091(0.043)	0.067(0.037)	0.980	24.012%	0.206(0.045)	0.184(0.045)	0.990	10.587%
II	**75**	0.060(0.029)	0.040(0.024)	1.000	32.846%	0.160(0.032)	0.138(0.035)	1.000	14.252%
II	**100**	0.049(0.022)	0.033(0.019)	1.000	32.636%	0.140(0.031)	0.122(0.030)	1.000	13.285%
III	**25**	0.451(0.236)	0.287(0.199)	0.960	35.967%	0.470(0.126)	0.380(0.125)	0.960	19.215%
III	**50**	0.270(0.121)	0.165(0.090)	0.990	36.812%	0.347(0.077)	0.280(0.072)	0.980	19.075%
III	**75**	0.211(0.091)	0.119(0.050)	0.990	39.786%	0.291(0.056)	0.229(0.050)	0.980	20.996%
III	**100**	0.171(0.076)	0.112(0.053)	1.000	32.405%	0.257(0.047)	0.213(0.043)	1.000	16.971%
IV	**25**	1.046(0.457)	0.804(1.070)	0.880	14.949%	0.727(0.177)	0.623(0.264)	0.860	12.086%
IV	**50**	0.728(0.246)	0.471(0.265)	0.960	35.859%	0.571(0.113)	0.469(0.146)	0.880	18.295%
IV	**75**	0.605(0.191)	0.384(0.154)	0.990	35.084%	0.497(0.079)	0.403(0.079)	0.960	18.539%
IV	**100**	0.462(0.162)	0.308(0.114)	1.000	30.776%	0.418(0.068)	0.342(0.064)	0.990	17.912%
V	**25**	1.896(0.766)	1.009(0.563)	0.980	44.803%	0.984(0.224)	0.703(0.211)	0.980	28.043%
V	**50**	1.396(0.478)	0.922(0.566)	0.900	32.353%	0.801(0.140)	0.648(0.218)	0.870	18.492%
V	**75**	1.057(0.303)	0.750(0.315)	0.950	28.473%	0.673(0.107)	0.567(0.152)	0.880	15.677%
V	**100**	0.914(0.261)	0.624(0.211)	0.980	29.296%	0.613(0.090)	0.502(0.087)	0.970	17.543%

6. Discussion

In this section, we briefly discuss the main results of this paper and how they can be interpreted from the perspective of previous studies, mainly those based on Data Envelopment Analysis. Our findings and their implications are also discussed. Some limitations of our approach are highlighted.

In this paper, we have introduced a new way of estimating production frontiers in engineering and microeconomics, which is based upon the same fundamentals of Support Vector Machines (SVM), which is a well-known machine learning technique. Our numerical results have demonstrated that the frontier estimator derived from the new methodology (DEAM) is better than that associated with Data Envelopment Analysis (DEA), which represents the standard non-parametric technique for determining technical efficiency in the literature. The bias and mean squared error obtained for DEAM are smaller in all the scenarios analyzed, regardless of the number of variables and DMUs.

In comparison with the standard literature, the new methodology is more flexible. It generates production possibility sets that satisfy convexity, free disposability in inputs and outputs, and contain all the observations, but they do not meet the postulate of minimal extrapolation. In contrast, DEA satisfies all the above properties. In particular, minimal extrapolation is the reason why DEA can be seen as an overfitted model to estimate the underlying Data Generating Process (DGP) that is behind the generation of the data sample. DEAM does not suffer from this overfitting problem. However, it is not evident where the production possibility set, estimated by a non-overfitted model, should be located in the input–output space to correctly approximate the underlying technology, which, by definition, is unknown to us. In this regard, in this paper, we have implemented for the first time a strategy based on the idea of Structural Risk Minimization (Vapnik [1]) and cross validation, introducing a new PAC (Probably Approximately Correct) bound in production theory with the aim of solving the overfitting problem linked to DEA.

Some other authors have tried to modify the standard DEA technique such that the new approaches work as inferential methods (with the focus on the DGP) rather than as mere descriptive tools. For example, Banker and Maindiratta [8] and Banker [9] associated

DEA with maximum likelihood. Simar and Wilson [10–12] adapted bootstrapping to DEA. Kuosmanen and Johnson [13,14] introduced the Corrected Concave Nonparametric Least Squares. Unfortunately, despite the importance of machine learning techniques in the current literature, there have been few attempts to adapt DEA to the field of machine learning (see, for example, Esteve et al. [7], or Olesen and Ruggiero [15]). In this sense, DEAM has allowed us to build a new bridge between these two worlds: machine learning and efficiency measurement.

Finally, we would like to highlight a clear limitation associated with the new approach. DEAM is linked to an intensive computational procedure based on cross validation. This feature contrasts sharply with the simplicity of Data Envelopment Analysis.

7. Conclusions and Future Work

In this paper, for the first time, a bound on the generalization error for a piece-wise linear hypothesis has been established in the context of Support Vector Regression (SVR), by also considering typical axioms from production theory: convexity and free disposability. It shapes a new nexus between non-parametric frontier analysis and machine learning in the line recently followed by Esteve et al. [7], Valero-Carreras et al. [32], and Olesen and Ruggiero [15]. The new formulation on the bound of the generalization error of this kind of hypothesis gives rise to a new way of bounding the whole expected error when we approximate a target function through a piece-wise linear function, also controlling the empirical error. Minimizing this bound led to the definition of a new model, called Data Envelopment Analysis-based Machines (DEAM), which generates production function estimations that seek a balance between the empirical error and the generalization error.

Classical non-parametric techniques, such as DEA, suffer from the overfitting problem because they assume the axiom of minimal extrapolation (Banker et al. [6], Afriat [28], and Farrell [33]). The DEAM model, however, is more flexible when it comes to estimating production frontiers through a cross-validation process, disregarding the minimal extrapolation axiom, as was shown by a computational experience in this paper.

Finally, we finish by mentioning several lines that pose interesting avenues for further research. The first one is the possibility of extending the method to model multi-output situations. This could be interesting for dealing with more realistic production situations, considering information on the correlation among several outputs. Second, we could use other transformation functions (kernel methods) for the input space, in the same way as standard Support Vector Regression.

Author Contributions: Conceptualization, N.M.G. and J.A.; methodology, N.M.G. and J.A.; software, D.V.-C.; validation, N.M.G. and D.V.-C.; formal analysis, N.M.G., J.A. and D.V.-C.; investigation, N.M.G., J.A. and D.V.-C.; resources, N.M.G., J.A. and D.V.-C.; data curation, N.M.G. and D.V.-C.; writing—original draft preparation, N.M.G., J.A. and D.V.-C.; writing—review and editing, N.M.G., J.A. and D.V.-C.; visualization, N.M.G. and D.V.-C.; supervision, J.A.; project administration, J.A.; funding acquisition, J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033 grant number PID2019-105952GB-I00, by Generalitat Valenciana grant number ACIF/2020/155, and by Miguel Hernández University of Elche grant number 01623/2020.

Acknowledgments: The authors are grateful to the two anonymous reviewers for providing constructive comments and helping in improving the contents and presentation of this paper. Additionally, the authors are thankful for grant PID2019-105952GB-I00 funded by Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033. D. Valero-Carreras is thankful for the financial support from the Generalitat Valenciana under grant ACIF/2020/155. Finally, N. Guerrero is thankful for the financial support from the Miguel Hernández University of Elche under grant 01623/2020.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Lemma 3. Let $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ be γ-shattered by

$$F = \begin{array}{l} \Big\{ x \mapsto \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} : \|\boldsymbol{x}\| \leq R, \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \geq 0, \\ \frac{\boldsymbol{w}_{p_x}}{\delta_{p_x}}\boldsymbol{x} + \frac{\beta_{p_x}}{\delta_{p_x}} \leq \frac{\boldsymbol{w}_p}{\delta_p}\boldsymbol{x} + \frac{\beta_p}{\delta_p}, \forall p_x, p \in \{1, \ldots, n\} \Big\} \end{array}$$

witnessed by $r_1, \ldots, r_d \in R$. Then, for all $b = (b_1, \ldots, b_d) \in \{-1, 1\}^d$, there are $\left\{\left(\boldsymbol{w}_p\right)_b\right\}_{p\in\{1,\ldots,n\}}$, $\left\{\left(\delta_p\right)_b\right\}_{p\in\{1,\ldots,n\}}$ and $\left\{\left(\beta_p\right)_b\right\}_{p\in\{1,\ldots,n\}}$ satisfying for all $i \in \{1, \ldots, d\}$ the following inequality:

$$b_i\left[\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b x_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i\right] \geq \gamma$$

Let set $S_0 \subset S$, and consider two cases:

- Case 1: If $\sum\{r_i : \boldsymbol{x}_i \in S_0\} \geq \sum\{r_i : \boldsymbol{x}_i \in S - S_0\}$, then $b_i = 1$ if and only if $\boldsymbol{x}_i \in S_0$
- Case 2: If $\sum\{r_i : \boldsymbol{x}_i \in S_0\} < \sum\{r_i : \boldsymbol{x}_i \in S - S_0\}$, then $b_i = 1$ if and only if $\boldsymbol{x}_i \in S - S_0$

Let us suppose that $\sum\{r_i : \boldsymbol{x}_i \in S_0\} \geq \sum\{r_i : \boldsymbol{x}_i \in S - S_0\}$, with $b_i = 1$ if and only if $\boldsymbol{x}_i \in S_0$ (*CASE 1*). For all $\boldsymbol{x}_i \in S_0$, we have

$$\begin{aligned} b_i\left[\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i\right] &= 1 \cdot \left[\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i\right] \\ &= \left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i \underset{x_i \in S_0 \subset S}{\geq} \gamma, \end{aligned}$$

that is

$$\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \geq r_i + \gamma.$$

Then, taking the sum over the elements in the set S_0, we obtain the expression $\sum_{i/x_i\in S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) \geq \sum_{i/x_i\in S_0}(r_i + \gamma)$; which yields the following inequality:

$$\sum_{i/x_i\in S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) \geq \sum\{r_i : x_i \in S_0\} + |S_0|\gamma.$$

From F, we have $\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \leq \left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_p}{\delta_p}\right)_b$, for all $p \in \{1, \ldots, P\}$. Thereby, the inequality

$$\left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b \underbrace{\sum_{i/x_i\in S_0} \boldsymbol{x}_i}_{\sum S_0} + |S_0|\left(\frac{\beta_p}{\delta_p}\right)_b = \sum_{i/x_i\in S_0}\left(\left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_p}{\delta_p}\right)_b\right) \geq \sum_{i/x_i\in S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right)$$

is satisfied $\forall p \in \{1, \ldots, n\}$. Finally,

$$\left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b \sum S_0 + |S_0|\left(\frac{\beta_p}{\delta_p}\right)_b \geq \sum\{r_i : \boldsymbol{x}_i \in S_0\} + |S_0|\gamma. \tag{A1}$$

Now, let $x_i \in S - S_0$, then

$$b_i\left[\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i\right] = (-1)\cdot\left[\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) - r_i\right] =$$

$$-\left(\left(\frac{w_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b x_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) + r_i \underset{\boldsymbol{x}_i \in S-S_0 \subset S}{\geq} \gamma$$

Then,

$$\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \leq r_i - \gamma.$$

Following the idea of applying the summary of elements, but now considering $x_i \in S - S_0$, the inequality $\sum_{i/x_i \in S-S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) \leq \sum_{i/\boldsymbol{x}_i \in S-S_0}(r_i - \gamma)$ holds. It can be rewritten as

$$\sum_{i/x_i \in S-S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right) \leq \sum\{r_i : \boldsymbol{x}_i \in S - S_0\} - |S - S_0|\gamma. \quad \text{(A2)}$$

Now, there is $x_{i'} \in S - S_0$ such that $\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{\boldsymbol{x}_{i'}}}}\right)_b x_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \leq \left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b$ for all $x_i \in S - S_0$. Consequently, we have that $|S - S_0|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \boldsymbol{x}_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \leq \sum_{i/x_i \in S-S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right)$. $|S - S_0| \geq 0$ because $|S - S_0|$ is the cardinal of the set $S - S_0$. Conversely, $\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \boldsymbol{x}_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \geq 0$ by definition of F. Then, $-|S - S_0|\left(\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \boldsymbol{x}_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right) \leq |S - S_0|\left(\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{\boldsymbol{x}_{i'}}}}\right)_b x_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right)$. Now, we can guarantee that

$$-|S - S_0|\left(\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{\boldsymbol{x}_{i'}}}}\right)_b x_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right) \leq \sum_{i/x_i \in S-S_0}\left(\left(\frac{\boldsymbol{w}_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b \boldsymbol{x}_i + \left(\frac{\beta_{p_{x_i}}}{\delta_{p_{x_i}}}\right)_b\right), \quad \text{(A3)}$$

and by (A2), the following inequality holds:

$$-|S - S_0|\left(\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \boldsymbol{x}_{i'} + \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right) \leq \sum\{r_i : \boldsymbol{x}_i \in S - S_0\} - |S - S_0|\gamma. \quad \text{(A4)}$$

Considering inequalities (25) and (28), for $p_{x_{i'}} \in \{1, \ldots, n\}$, we have that $\left(\begin{array}{c} A \geq B \\ C \leq D \end{array}\right\} \Rightarrow A - C \geq B - D$.)

$$\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \sum S_0 + |S_0|\left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b + |S - S_0|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \boldsymbol{x}_{i'} + |S - S_0|\left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \geq$$
$$\sum\{r_i : \boldsymbol{x}_i \in S_0\} + |S_0|\gamma - \sum\{r_i : \boldsymbol{x}_i \in S - S_0\} + |S - S_0|\gamma$$

and then,

$$\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b [\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}] + |S|\left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \geq$$

$$\sum\{r_i : \boldsymbol{x}_i \in S_0\} - \sum\{r_i : \boldsymbol{x}_i \in S - S_0\} + |S|\gamma.$$

Under the supposition in the case 1 that $\sum\{r_i : \boldsymbol{x}_i \in S_0\} \geq \sum\{r_i : \boldsymbol{x}_i \in S - S_0\}$, we have that

$$\sum\{r_i : \boldsymbol{x}_i \in S_0\} - \sum\{r_i : \boldsymbol{x}_i \in S - S_0\} \geq 0,$$

which implies that

$$\left(\sum\{r_i : \boldsymbol{x}_i \in S_0\} - \sum\{r_i : \boldsymbol{x}_i \in S - S_0\}\right) + |S|\gamma \geq |S|\gamma.$$

Therefore, for $p_{x_{i'}} \in \{1, \ldots, n\}$, we have

$$\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b [\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}] + |S|\left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \geq |S|\gamma,$$

that is,

$$\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b [\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}] \geq |S|\left(\gamma - \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right). \tag{A5}$$

Under the Cauchy–Schwarz inequality, we have

$$\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b [\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}] \leq \left\|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right\| \|\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}\|,$$

and then, we have

$$\|\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}\| = \|\sum S_0 - \sum(S - S_0) + \sum(S - S_0) + |S - S_0|\boldsymbol{x}_{i'}\| \underset{Triangular}{\leq} \|\sum S_0 - \sum(S - S_0)\| +$$
$$+\|\sum(S - S_0)\| + |S - S_0|\|\boldsymbol{x}_{i'}\| \underset{Triangular}{\leq} \|\sum S_0 - \sum(S - S_0)\| + \sum_{i/x_i \in S - S_0} \|\boldsymbol{x}_i\| + |S - S_0|\|\boldsymbol{x}_{i'}\| \underset{\|\boldsymbol{x}_i\| \leq r, \forall i \in \{1,\ldots,n\}}{\leq}$$
$$\|\sum S_0 - \sum(S - S_0)\| + |S - S_0|r + |S - S_0|r = \|\sum S_0 - \sum(S - S_0)\| + 2|S - S_0|r \underset{|S - S_0| \leq |S|}{\leq}$$
$$\|\sum S_0 - \sum(S - S_0)\| + 2|S|r.$$

In this way, it is possible to obtain

$$\left\|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right\| (\|\sum S_0 - \sum(S - S_0)\| + 2|S|r) \geq \left\|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right\| \|\sum S_0 + |S - S_0|\boldsymbol{x}_{i'}\| \geq |S|\left(\gamma - \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right).$$

Because $\left\|\left(\frac{\boldsymbol{w}_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b\right\| \leq \underset{p \in \{1,\ldots,n\}}{Max} \left\|\left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b\right\|$ and $\gamma - \left(\frac{\beta_{p_{x_{i'}}}}{\delta_{p_{x_{i'}}}}\right)_b \geq \underset{p \in \{1,\ldots,n\}}{Min} \left(\gamma - \left(\frac{\beta_p}{\delta_p}\right)_b\right)$, then

$$\underset{p \in \{1,\ldots,n\}}{Max} \left\|\left(\frac{\boldsymbol{w}_p}{\delta_p}\right)_b\right\| (\|\sum S_0 - \sum(S - S_0)\| + 2|S|r) \geq |S| \underset{p \in \{1,\ldots,n\}}{Min} \left(\gamma - \left(\frac{\beta_p}{\delta_p}\right)_b\right).$$

Finally,

$$\left\|\sum S_0 - \sum(S - S_0)\right\| \geq |S| \left(\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left(\gamma - \left(\frac{\beta_p}{\delta_p}\right)_b\right)}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\left(\frac{w_p}{\delta_p}\right)_b\right\|} - 2r \right), \ \forall S_0 \subset S.$$

The proof for case 2 is analogous. □

Proof of Theorem 3. By Lemma 3, we have

$$\left\|\sum S_0 - \sum(S - S_0)\right\| \geq |S| \left(\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left\{\gamma - \frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\frac{w_p}{\delta_p}\right\|} - 2r \right),$$

for every subset $S_0 \subseteq S$, with $S = \{x_1, \ldots, x_d\}$ being an input learning sample γ-shattered through F defined in (10). Additionally, by Lemma 2, for all $S \subseteq R_+^m$ with $\|x\| \leq r$ for $x \in S$, some $S_0 \subseteq S$ satisfies the following condition:

$$\left\|\sum S_0 - \sum(S - S_0)\right\| \leq \sqrt{|S|}r.$$

Then, for certain $S_0 \subseteq S$, we have

$$\left\|\sum S_0 - \sum(S - S_0)\right\| \geq |S| \left(\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left\{\gamma - \frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\frac{w_p}{\delta_p}\right\|} - 2r \right) \text{ and } \left\|\sum S_0 - \sum(S - S_0)\right\| \leq \sqrt{|S|}r.$$

Therefore,

$$|S| \left(\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left\{\gamma - \frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\frac{w_p}{\delta_p}\right\|} - 2r \right) \leq \sqrt{|S|}r.$$

Finally,

$$|S| \leq \left(\frac{r}{\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left\{\gamma - \frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\frac{w_p}{\delta_p}\right\|} - 2r} \right)^2.$$

Because this is true for all S γ-shattered by F, it will be also true for the largest set γ-shattered by F, which means that $fat_F(\gamma)$ will be bound in that way:

$$fat_F(\gamma) \leq \left(\frac{r}{\frac{\underset{p\in\{1,\ldots,n\}}{Min}\left\{\gamma - \frac{\beta_p}{\delta_p}\right\}}{\underset{p\in\{1,\ldots,n\}}{Max}\left\|\frac{w_p}{\delta_p}\right\|} - 2r} \right)^2.$$

□

References

1. Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
2. Vapnik, V. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1992; pp. 831–838.
3. Blanco, V.; Puerto, J.; Salmerón, R. Locating hyperplanes to fitting set of points: A general framework. *Comput. Oper. Res.* **2018**, *95*, 172–193.
4. Blanco, V.; Puerto, J.; Rodriguez-Chia, A.M. On lp-Support Vector Machines and Multidimensional Kernels. *J. Mach. Learn. Res.* **2020**, *21*, 14.
5. Charnes, A.; Cooper, W.W.; Rhodes, E. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444. [CrossRef]
6. Banker, R.D.; Charnes, A.; Cooper, W.W. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manag. Sci.* **1984**, *30*, 1078–1092. [CrossRef]
7. Esteve, M.; Aparicio, J.; Rabasa, A.; Rodriguez-Sala, J.J. Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees. *Expert Syst. Appl.* **2020**, *162*, 113783. [CrossRef]
8. Banker, R.D.; Maindiratta, A. Maximum likelihood estimation of monotone and concave production frontiers. *J. Product. Anal.* **1992**, *3*, 401–415. [CrossRef]
9. Banker, R.D. Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Manag. Sci.* **1993**, *39*, 1265–1273. [CrossRef]
10. Simar, L.; Wilson, P.W. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Manag. Sci.* **1998**, *44*, 49–61.
11. Simar, L.; Wilson, P.W. A general methodology for bootstrapping in non-parametric frontier models. *J. Appl. Stat.* **2000**, *27*, 779–802.
12. Simar, L.; Wilson, P.W. Statistical inference in nonparametric frontier models: The state of the art. *J. Product. Anal.* **2000**, *13*, 49–78. [CrossRef]
13. Kuosmanen, T.; Johnson, A.L. Data envelopment analysis as nonparametric least-squares regression. *Oper. Res.* **2010**, *58*, 149–160. [CrossRef]
14. Kuosmanen, T.; Johnson, A. Modeling joint production of multiple outputs in StoNED: Directional distance function approach. *Eur. J. Oper. Res.* **2017**, *262*, 792–801.
15. Olesen, O.B.; Ruggiero, J. The hinging hyperplanes: An alternative nonparametric representation of a production function. *Eur. J. Oper. Res.* **2022**, *296*, 254–266.
16. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
17. Valiant, L.G. A theory of the learnable. *Commun. ACM* **1984**, *27*, 1134–1142.
18. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, MA, USA, 2000.
19. Bartlett, P.; Shawe-Taylor, J. *Generalization Performance of Support Vector Machines and Other Pattern Classifiers. Adv. Kernel Methods Support Vector Learn*; MIT Press: Cambridge, MA, USA, 1999; pp. 43–54.
20. Vazquez, E.; Walter, E. Multi-output suppport vector regression. *IFAC Proc. Vol.* **2003**, *36*, 1783–1788. [CrossRef]
21. Villa, G.; Lozano, S.; Redondo, S. Data envelopment analysis approach to energy-saving projects selection in an energy service company. *Mathematics* **2021**, *9*, 200. [CrossRef]
22. Sahoo, B.K.; Saleh, H.; Shafiee, M.; Tone, K.; Zhu, J. An Alternative Approach to Dealing with the Composition Approach for Series Network Production Processes. *Asia-Pac. J. Oper. Res. (APJOR)* **2021**, *38*, 2150004.
23. Amirteimoori, A.; Sahoo, B.K.; Charles, V.; Mehdizadeh, S. Stochastic Network Data Envelopment Analysis. In *Stochastic Benchmarking*; Springer: Cham, Switzerland, 2022; pp. 77–117.
24. Färe, R.; Primont, D. Distance functions. In *Multi-Output Production and Duality: Theory and Applications*; Springer: Dordrecht, The Netherlands, 1995; pp. 7–41.
25. Briec, W.; Lesourd, J.B. Metric distance function and profit: Some duality results. *J. Optim. Theory Appl.* **1999**, *101*, 15–33.
26. Cooper, W.W.; Seiford, L.M.; Tone, K. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*; Springer: New York, NY, USA, 2007; Volume 2.
27. Briec, W. Hölder distance function and measurement of technical efficiency. *J. Product. Anal.* **1999**, *11*, 111–131.
28. Afriat, S.N. Efficiency estimation of production functions. *Int. Econ. Rev.* **1972**, *13*, 568–598.
29. Mangasarian, O.L. Arbitrary-norm separating plane. *Oper. Res. Lett.* **1999**, *24*, 15–23.
30. Aparicio, J.; Pastor, J.T. A well-defined efficiency measure for dealing with closest targets in DEA. *Appl. Math. Comput.* **2013**, *219*, 9142–9154.
31. Charles, V.; Aparicio, J.; Zhu, J. The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *Eur. J. Oper. Res.* **2019**, *279*, 929–940.
32. Valero-Carreras, D.; Aparicio, J.; Guerrero, N.M. Support vector frontiers: A new approach for estimating production functions through support vector machines. *Omega* **2021**, *104*, 102490.
33. Farrell, M.J. The measurement of productive efficiency. *J. R. Stat. Soc. Ser. A* **1957**, *120*, 253–281.