European Journal of Operational Research 312 (2024) 684–699

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Estimating production functions through additive models based on regression splines

Victor J. España[a], Juan Aparicio[a,b,*], Xavier Barber[a], Miriam Esteve[a]

[a] Center of Operations Research (CIO). Miguel Hernandez University of Elche. Avda. de la universidad, s/n, 03202 Elche, Spain
[b] valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence. Camino de Vera, s/n, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 5 January 2022
Accepted 19 June 2023
Available online 23 June 2023

Keywords:
Data envelopment analysis
Additive models
Machine learning
Overfitting

ABSTRACT

This paper introduces a new methodology for the estimation of production functions satisfying some classical production theory axioms, such as monotonicity and concavity, which is based upon the adaptation of an additive version of the machine learning technique known as Multivariate Adaptive Regression Splines (MARS). The new approach shares the piece-wise linear shape of the estimator associated with Data Envelopment Analysis (DEA). However, the new technique is able to surmount the overfitting problems associated with DEA by resorting to generalized cross-validation. In this paper, a computational experience was employed to measure how well the new approach performs, showing that it can reduce the mean squared error and bias of the estimator of the true production function in comparison with DEA and the more recent Corrected Concave Non-Parametric Least Squares (C^2NLS) methodology. We also show that the success of the new approach depends on whether or not interactions among variables prevail and the degree of non-additivity of the true production function to be estimated.

1. Introduction

Data Envelopment Analysis (DEA) (Charnes et al., 1978, Banker et al., 1984) is one of the most widely used techniques for the estimation of production functions and efficiency measurement. DEA relies on the construction of a technology in the space of inputs and outputs that satisfies certain classical axioms of production theory (e.g., free disposability and convexity). It is a non-parametric data-driven approach with a lot of 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 existing methods (see, e.g., the Stochastic Frontier Analysis approach by Aigner et al., 1977). However, Data Envelopment Analysis has been criticized for its non-statistical nature, even being labeled as a pure descriptive tool for frontier sample data with little inferential power (exclusively based on the property of consistency) (Esteve et al., 2020). In fact, 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 (see Esteve et al., 2020, Tsionas, 2022, Valero-Carreras et al., 2022, and Molinos-Senante et al., 2023). In line with this, various authors have attempted to correct these deficiencies within the non-parametric approach over the last few decades, introducing complementary and alternative methodologies to DEA. For example, Simar and Wilson (1998, 2000a) adapted the bootstrapping methodology to the determination of confidence intervals for estimating efficiency scores obtained via DEA. Kuosmanen and Johnson (2010, 2017) introduced the Corrected Concave Non-parametric Least Squares (C^2NLS), whose objective is to provide a pointwise estimation of the theoretical production function that generated the observed data sample. However, the problem of estimating production functions and efficiency through Machine Learning (ML) techniques, taking advantage of their non-parametric and data driven features, has been relatively less addressed in the literature. This scarcity of bridges between machine learning and production function estimation is only justified by the novelty of the ML methods and the very recent interest they have aroused in all areas of science. However, we must highlight the contributions made by Esteve et al. (2020), Valero-Carreras et al. (2021) and Olesen and Ruggiero (2022) in this regard. The first authors defined Efficiency Analysis Trees with the objective of efficiency frontier estimation; largely built on the adaptation of the Classification and Regression Analysis Trees (CART) approach by Breiman et al. (1984). The second authors adapted the machine learning method known as Support Vector Regression by Drucker et al. (1997) to be used in the production function estimation setting, satisfying usual axioms in microeconomics. Lastly, the third authors introduced Breiman's Hinging Hyperplanes function approximation as

* Corresponding author.
E-mail address: j.aparicio@umh.es (J. Aparicio).

https://doi.org/10.1016/j.ejor.2023.06.035

a flexible estimator of production functions. Other related articles are Parmeter and Racine (2013), Daouia et al. (2016), Zhu et al. (2018), Zhu (2020), Dellnitz (2022) and Esteve et al. (2023).

Alternative methodologies devoted to solving the lack of robustness of the DEA technique are based on the estimation of quantile frontiers, instead of the estimation of full frontiers that envelops all the observations. In this context, contributions such as Aragon et al. (2005), Wang and Wang (2013) and Wang et al. (2014) must be considered. The first authors introduced the idea of using conditional quantiles of a suitable distribution linked to the production process for the construction of a non-parametric estimator of the efficient frontier. These ideas were extended in Daouia and Simar (2007). The second contribution presents a non-parametric smooth multivariate estimation based on kernel quantile regression with shape constraints: non-decreasing monotony and concavity. Finally, the third authors introduce a non-parametric shape-restricted quantile regression methodology in a two-step approach. First, they identify the fitted values that minimize a loss criterion imposing non-decreasing monotonicity and concavity restrictions; and secondly, they build a non-decreasing monotonic and concave estimator of the target function.

This paper is in line with Esteve et al. (2020), Valero-Carreras et al. (2021) and Olesen and Ruggiero (2022) and its main objective is to approximate the estimation of production functions to the field of Machine Learning. With this purpose in mind, this article shows how an additive version of the technique known as Multivariate Adaptive Regression Splines (MARS) by Friedman (1991), is adapted for the first time in the literature to be used for the estimation of production functions. MARS is a non-parametric splines-based method that extends linear regression models by including nonlinearities and interactions between predictors. This technique is useful to approximate a target function based on piecewise polynomials. To do this, the predictors' domain is divided into a certain number of intervals. The point in the predictor space that splits two of these intervals and that typically identifies a trend change in the data patterns is commonly known as a knot. Precisely, the performance of spline-based methods can be limited due to the need to determine, a priori, the position and number of knots, a task that can be challenging in high-dimensionality scenarios (Friedman et. al, 2001). To overcome this weakness, MARS applies a recursive partitioning algorithm through an adaptive process that achieves an optimal selection of the location of each knot. In particular, MARS is grounded on two automatic processes, implemented as algorithms. The first one is a forward selection process, which splits the predictor space recursively into (non-necessarily disjoint) subspaces based on an intensive search of knots throughout the range of the predictors. These knots are used to define a set of transformation functions (called basis functions) of the original predictors through splines. The second process is a backward removing mechanism. At each forward step, the spline function that minimizes the training error is added as a new term of the model. Once the set of possible basis functions has been defined or the error has not been sufficiently reduced, the backward algorithm sequentially removes those terms that will achieve least degradation of the model performance. MARS avoids the problem of data overfitting in this way.

After MARS was introduced by Friedman, various authors have suggested modifications to the method to address possible limitations or to achieve additional properties. Chen et al. (1999) presented a quintic function for smoothing the estimator and thus obtain a MARS model with continuous second derivatives. Bakin et al. (2000) developed a new version of MARS, called BMARS, using second-order B-splines instead of truncated linear functions with the aim of obtaining numerical stability. Tsai and Chen (2005) explored two new variants of MARS: first, applying automatic stopping rules based on the (adjusted) coefficient of determination instead of allowing the forward algorithm to grow until the maximum number of basis functions is reached (deleting the backward step); and, secondly, developing a robust version to decrease the order of the interaction terms. In this way, Tsai and Chen (2005) managed to reduce the computational cost of MARS and improve its performance against extreme values. Taylan et al. (2010) provided parameter estimates for generalized partial linear models with B-splines using conic quadratic programming that may serve as a basis for further research into MARS. Weber et al. (2011) suggested a new approach, called CMARS, where the backward stepwise algorithm is modified by using a penalized residual sum of squares, as a Tikhonov regularization problem, which can be expressed as a conic quadratic programming problem. Later, Özmen et al. (2011) and Özmen and Weber (2014) enhanced CMARS (RCMARS) and MARS (RMARS), respectively, by robust optimization techniques to deal with data uncertainty (see also Özmen et al., 2017). A further improvement of MARS on the already formulated CMARS was developed by Yazici et al. (2015), who included the bootstrap method (BCMARS) to obtain an empirical distribution of the fitted parameters to determine their significance. Koc and Bozdogan (2015) presented another alternative to the conventional backward algorithm by using the information-theoretic measure of complexity (ICOMP) for model selection. Martinez et al. (2015) provided a convex version of MARS by altering the form of introducing interaction terms and constraining the coefficients to eliminate the inherent non-convexity. Additionally, Zhang (1994) and Koc and Iyigun (2014) modified the forward algorithm using new knot selection procedures. Finally, Murat (2021) proposed a strategy to detect outliers via the variable selection process in MARS. To do that, a designed matrix is built by adding as many dummy variables to the observed data as potential outliers are considered.

In this paper, we introduce an additive version of MARS to estimate production functions. Shape-restricted additive regression belongs to the literature devoted to additive models in Statistics. Some interesting contributions in this line are: Bacchetti (1989), who developed additive isotonic (monotonic) multivariate models using an iterative application of the pool-adjacent-violators algorithm (Ayer et al. 1955); Chen and Samworth (2016), who proposed a general additive model imposing monotony and/or curvature constraints on each component of the additive function; Mammen and Yu (2007), who presented a backfitting algorithm based on iterative applications of least squares isotone to each additive component; and Meyer (2013), who proposed a more general semiparametric additive constrained regression. In particular, in our production context, the additive version of MARS that we propose in this paper requires the fulfilment of classical postulates in microeconomics within production theory. Specifically, we refer to the monotonicity and concavity properties of the production function. These conditions represent shape constraints that must be considered when proposing a suitable estimator of the target function. The estimator yielded by Data Envelopment Analysis is easily determined by Linear Programming in just one step and satisfies the previously mentioned shape constraints. Furthermore, our estimator will be a piecewise linear function, as happens with the estimator determined through DEA. Accordingly, the technologies estimated through the new approach will include the technology obtained through DEA as a subset; since DEA satisfies the minimal extrapolation principle (i.e., it fits the data sample as closely as possible). It is precisely this principle that causes the overfitting problem occurring in DEA.

The contributions of this paper are two-fold. First, we introduce Friedman's MARS in the production framework and show that this technique is important for non-parametric production function estimation. To do that, we adapt an additive version of MARS forward and backward algorithms for estimating monotone and con-

cave target functions that, additionally, must envelop the data from above. In particular, we prove that the yielded production function estimator satisfies all these desired properties. Secondly, we check the validity of the new approach in comparison with standard DEA and the recent Corrected Concave Non-Parametric Least Squares (C^2NLS) by Kuosmanen and Johnson (2010, 2017) through a simulation experience with six different scenarios. We will show that the new technique performs better than DEA in almost all the cases studied. Moreover, when the number of considered inputs is increased, the percentage of improvement is even higher. Regarding the comparison with respect to Corrected Concave Non-parametric Least Squares, the new technique outperforms C^2NLS in all the scenarios considered except one. Nevertheless, the success of the new approach depends on whether or not there are interaction terms among variables as well as the degree of non-additivity of the true production function to be estimated. At this point, it is worth mentioning that Vidoli (2011) also resorted to MARS in a context of production efficiency measurement. From a methodological perspective, Vidoli (2011) introduced an approach that uses two stages. At the first stage, it estimates a conditional robust production function, by following the conditional order-m approach. At the second stage, the standard MARS model is estimated on the frontier identified by units that present values of efficiency greater than or equal to 1 when the model of the first stage is used. Consequently, in that paper, the attention is paid to the evaluation of the effects of external variables Z and, additionally, the standard MARS model is directly applied. In contrast, our approach is very different. First, we do not focus our attention on Z variables. Second, we do not combine methods previously introduced in the literature to generate a new one. We tailor MARS to estimate production functions. This means that we force the output predictor to satisfy certain microeconomic postulates (shape constraints). In particular, envelopmentness, concavity and non-decreasing monotonicity. To do that, we add certain new constraints to the optimization model used in each step of the standard MARS algorithm. Third, we carried out a complete simulation experiment to show the superiority of our approach in comparison with DEA and Corrected Concave Non-Parametric Least Squares, while Vidoli (2011) did not simulate and directly applied the approach to an empirical database.

The paper is organized as follows. Section 2 introduces the background of the paper. Section 3 shows how an additive version of additive MARS has been adapted to provide suitable estimations of production functions in microeconomics. Section 4 employs computational experiments with simulated data to corroborate how well the new approach performs. Finally, Section 5 presents our conclusions.

2. Background

This section offers an overview of key concepts related to Data Envelopment Analysis and Multivariate Adaptive Regression Splines. We will also introduce some notation.

2.1. Data envelopment analysis (DEA)

Let us consider n units, whose technical efficiency level needs to be evaluated. These units (firms, organizations, etc.), called Decision Making Units (DMUs), consume $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{mi}) \in R_+^m$ inputs (i.e., resources) to produce $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{si}) \in R_+^s$ outputs (i.e., goods or services)[1]. The relative efficiency of each DMU comprising the sample is assessed with respect to the so-called production possibility set or technology, which encompasses the set of all combinations that potentially are technically feasible $(\boldsymbol{x}, \boldsymbol{y})$ and, broadly speaking, can be expressed as follows:

$$\varphi = \left\{(\boldsymbol{x}, \boldsymbol{y}) \in R_+^{m+s} : \boldsymbol{x} \text{ can produce } \boldsymbol{y}\right\} \tag{1}$$

Certain assumptions are usually made on this set, such as free disposability of inputs and outputs; meaning that if $(\boldsymbol{x}, \boldsymbol{y}) \in \varphi$, then $(\boldsymbol{x}', \boldsymbol{y}') \in \varphi$, as long as $\boldsymbol{x}' \geq \boldsymbol{x}$ and $\boldsymbol{y}' \leq \boldsymbol{y}$, and convexity; which implies that if $(\boldsymbol{x}, \boldsymbol{y}) \in \varphi$ and $(\boldsymbol{x}', \boldsymbol{y}') \in \varphi$, then $\lambda(\boldsymbol{x}, \boldsymbol{y}) + (1-\lambda)(\boldsymbol{x}', \boldsymbol{y}') \in \varphi, \forall \lambda \in [0, 1]$ (see Färe and Lovell, 1978). Deterministicness is another typical assumption made about these sets (see Banker et al., 1984), which guaranties that $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \varphi$, $\forall i = 1, \ldots, n$. In other words, the last axiom states that the production possibility set contains all the DMUs that belong to the data sample, and, graphically, its boundary envelops the observed data cloud from above.

Particularly, when $s = 1$, this framework is restricted to the key notion of production functions, defined as the maximum producible output for a given input profile. Also, the free disposability assumption, known in this case as monotonicity, implies that f is a monotone non-decreasing function, that is, if $\boldsymbol{x} \leq \boldsymbol{x}'$ then $f(\boldsymbol{x}) \leq f(\boldsymbol{x}')$. Accordingly, the technology is defined as:

$$\varphi = \left\{(\boldsymbol{x}, y) \in R_+^{m+1} : y \leq f(\boldsymbol{x})\right\}. \tag{2}$$

Hereinafter, we will turn our attention to estimating production functions. The existing methodologies for this purpose are either parametric or non-parametric. Some of the advantages of the non-parametric approach are the non-imposition of a prior functional form on the underlying technology (e.g., a Cobb-Douglas production function) and its ability to deal naturally with multi-output scenarios without assigning prior weights to the inputs and outputs. In contrast, the non-parametric approaches also have some drawbacks in comparison with their parametric counterparts: low robustness to outliers, they are closely related to the problem of overfitting, or they do not consider random error when it comes to inefficiency measurements.

Among the non-parametric techniques, Data Envelopment Analysis (DEA) is one of the most applied methods in practice. DEA (Charnes et al. 1978, Banker et al. 1984) is a non-parametric methodology for estimating the efficient frontier of φ by means of the satisfaction of certain postulates: free disposability, convexity, deterministicness and minimal extrapolation. The principle of minimal extrapolation is an additional requirement for selecting the most conservative estimator that satisfies free disposability and convexity and that contains the observed data.

Banker et al. (1984) put forward the DEA estimator of the production possibility set φ in the following way:

$$\varphi_{DEA} = \left\{ \begin{array}{l} (\boldsymbol{x}, \boldsymbol{y}) \in R_+^{m+s} : y_r \leq \sum_{i=1}^{n} \lambda_i y_{ri}, \forall r, \\ \quad x_j \geq \sum_{i=1}^{n} \lambda_i x_{ji}, \forall j, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, \forall i \end{array} \right\}. \tag{3}$$

Next, we present a graphical example of the DEA estimator of a production function (see Fig. 1). We can observe that DEA is built in a piece-wise linear manner and its corresponding estimator is

[1] Bold face letters, as $\boldsymbol{x}$, $\boldsymbol{y}$ or $\boldsymbol{\alpha}$, will denote vectors throughout the manuscript. Furthermore, non-bold face letters, as λ, will denote scalars. Additionally, the non-bold face letter x_j, with only one subscript, will denote the j-th (input) variable, while the non-bold face letter y_r, with only one subscript, will denote the r-th (output) variable. The non-bold face letter x_{ji}, with two subscripts, will denote the j-th input value of the i-th observation in the sample, whereas the non-bold face letter y_{ri}, with two subscripts, will denote the r-th output value of the i-th observation in the sample. The way of denoting the input-output vector associated with the i-th observation will be $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, by resorting to bold face letters and only one subscript. In the case of Greek letters as the vector $\boldsymbol{\alpha}$, which will be associated with parameters, each component will be denoted by using non-bold face letters and only a subscript: α_p.

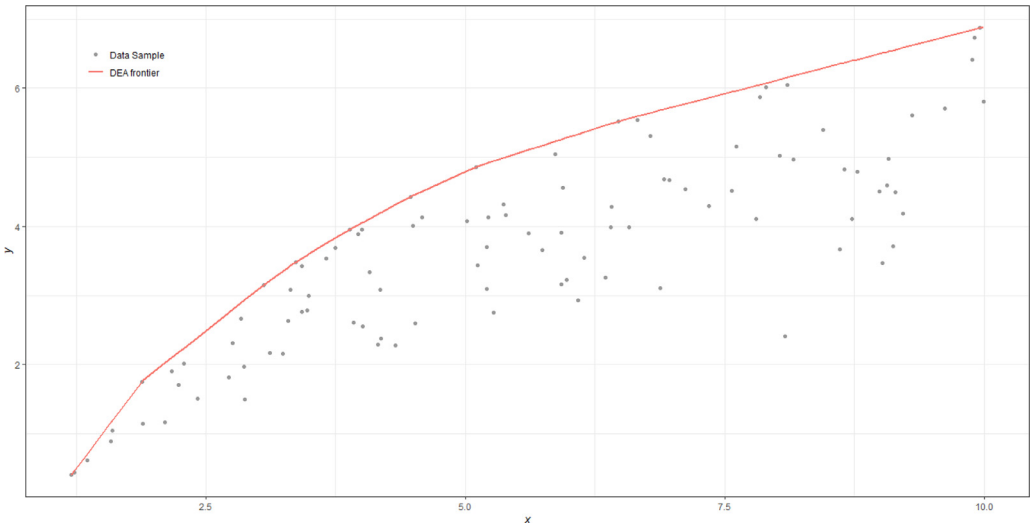

Fig. 1. An example of a DEA estimate.

monotonically non-decreasing. Furthermore, convexity of the production possibility set implies concavity. Finally, the application of the minimal extrapolation principle forces DEA to overfit the data sample. Consequently, it can effectively describe the observed data from an efficiency evaluation perspective, but it is not able to furnish an adequate generalization, i.e., a good evaluation of the actual production function that is behind the data generation[2].

2.2. Multivariate adaptive regression splines (MARS)

Contrary to the estimation of efficient frontiers, which is based on the study of extreme behaviors, traditional regression techniques in statistics seek to explain or predict mean behaviors. Regression analysis lies in modeling the dependence of a response variable y and a set of predictor variables $x_1, \ldots, x_m$ from a data sample perturbed by noise. Thus, the underlying data structure can be described by the following expression:

$$y = f(x_1, \ldots, x_m) + \varepsilon. \tag{4}$$

Here, the first term $f(x_1, \ldots, x_m)$ captures the relationship between the response variable y and the set of selected predictors $\boldsymbol{x} = (x_1, \ldots, x_m)$, whereas ε reflects the variability in y that cannot be explained from the selected predictors. Then, the goal here is to estimate a mathematical expression $\widehat{f}(x_1, \ldots, x_m)$ that can approximate the target function $f(x_1, \ldots, x_m)$ as much as possible. The different methods applied for this purpose can be classified as parametric and non-parametric.

The most recognized parametric technique of regression analysis is Linear Regression. Under its parametric condition, f is assumed to be a linear combination between the response variable and the predictors: $f(\boldsymbol{x}) = \sum_{j=1}^{m} \gamma_j x_j$. The simplicity of this approach justifies its widespread use in social science disciplines. Nevertheless, this simplicity is a direct consequence of the restrictive assumptions imposed on the estimated function (e.g., linear dependency, homoscedasticity, etc.), which in many cases result in very poor fits. In contrast, non-parametric procedures reject the prior assumptions made about the probability distributions of the data and the relationships between them. Consequently, the predictor function is made more flexible. From this perspective, spline-based methods stand out since they strive to approximate a function f based on piecewise polynomials. To do this, the domain $(x_1, \ldots, x_m) \in D \subset R^m$ is divided into $K-1$ contiguous intervals by K points and a polynomial is estimated in each interval only from the samples contained therein. The points in the input space that divide two contiguous intervals and that typically identify a trend change in the data patterns are commonly known as knots. Precisely, the performance of spline-based methods can be limited due to the need to determine, a priori, the position and number of knots, a task that can be challenging in high-dimensionality scenarios (Friedman et. al, 2001). To overcome the aforementioned weaknesses, there are some techniques that follow this methodology built upon the recursive partitioning technique through an adaptive process that achieves an optimal selection of the knot locations in a data-driven approach. Here, we can mention Classification and Regression Trees (CART) by Breiman et al. (1984) and Multivariate Adaptive Regression Splines (MARS) by Friedman (1991) as two relevant non-parametric techniques. While CART estimates step functions, MARS fits functions with different gradients in each interval. See Zhang and Singer (2010) for a comparison between both techniques.

In particular, MARS is a non-parametric regression technique especially designed to deal with high-dimensional scenarios with a non-linear relationship and complex interactions in the data. The resulting model is continuous with continuous first derivatives and is constructed as a linear combination of splines or product of splines. MARS can be seen as an extension of the CART technique.

The model-fitting process in MARS consists of two stepwise procedures: a forward selection and a backward elimination. The forward selection divides the input space recursively into new subspaces, not necessarily disjointly, based on an intensive search of knots along the range of the predictors. These knots are used to make up a set of transformation functions on the original predictors (basis functions) through splines. At each forward step, the spline function that most reduces the training error is added as a new term of the model. Once the number of basis functions preset by the user has been created, or the error is not sufficiently reduced, the backward algorithm sequentially deletes those terms whose removal implies the least degradation of the model performance. In this manner, overfitting is avoided and an assessment and selection of predictors is made.

The approximation function can be expressed as:

$$\hat{f}_B(\boldsymbol{x}; \boldsymbol{\gamma}(B)) = \sum_{b=1}^{|B|} \gamma_b(B) \cdot B_b(\boldsymbol{x}), \tag{5}$$

where B is the set of terms or basis functions that the model contains, $B_b(\boldsymbol{x}) \in B$ is a transformation on the original predictors

[2] Data Envelopment Analysis utilizes statistical consistency for approximating the underlying production function (Simar and Wilson, 2000b), i.e., the quality of the approximation is exclusively limited to the sample size.

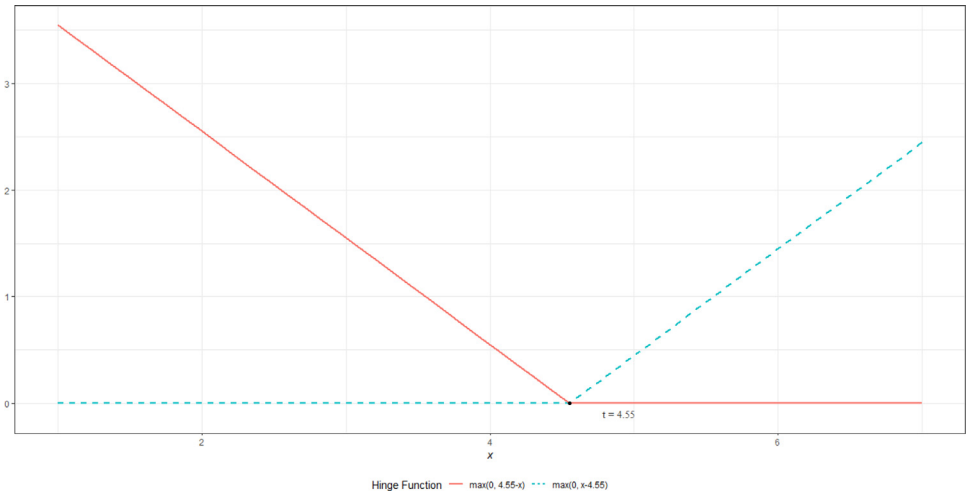

Fig. 2. An example of hinge functions in MARS.

$\boldsymbol{x} = (x_1, \ldots, x_m)$, $|B|$ is the cardinal of set B and $\boldsymbol{\gamma}(B) = (\gamma_1(B), \ldots, \gamma_{|B|}(B))$ is a vector of unknown coefficients to be estimated.

To create the set B during the forward algorithm, Friedman (1991) proposes implementing a strategy based on selecting a two-sided truncated univariate spline of degree 1 as a basis function:

$$b^{\pm}(x_j - t_j) = \left[\pm(x_j - t_j)\right]_+. \tag{6}$$

The previous expression can also be described by:

$$b^{+}(x_j - t_j) = \left[x_j - t_j\right]_+ = \max\left(0, x_j - t_j\right), \text{ and}$$
$$b^{-}(x_j - t_j) = \left[t_j - x_j\right]_+ = \max\left(0, t_j - x_j\right). \tag{7}$$

These piecewise linear basis functions can also be called reflected pairs or pairs of hinge functions. The main idea in MARS is to create reflected pairs by searching over all combination of predictors x_j, $j = 1, \ldots, m$, and all observed values of that predictors $x_{ji}, i = 1, \ldots, n$, as a candidate knot. Therefore, the collection of reflected pairs is:

$$\aleph = \left\{ \left\{ (x_j - t_j)_+, (t_j - x_j)_+ \right\} | t_j \in \left\{x_{j1}, x_{j2}, \ldots, x_{jn}\right\}, j = 1, \ldots, m \right\}. \tag{8}$$

Next, we show an example of a pair of hinge functions. Fig. 2 shows a knot t_j in $x_1 = 4.55$. In this manner, a reflected pair is created from the following expressions: $(4.55 - x_1)_+$ and $(x_1 - 4.55)_+$. The former, the left-side hinge function, is canceled under the condition that $x_1 \geq 4.55$ and it has a negative slope in the left-side. Conversely, the right-side hinge function is canceled for the data that satisfies the condition $x_1 \leq 4.55$ and it has a positive slope in the right-side.

The algorithm is initialized with $B_1(\boldsymbol{x}) = 1$ to set the initial region to the entire domain. Next, we select the pair of hinge functions from (8), multiplied by another basis function already entered in the model (parent term), that most reduce the mean of the residual sum of squares in the training sample (the lack-of-fit criterion). In the case of considering a single response variable (y), the criterion is defined as follows:

$$LOF = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_B(\boldsymbol{x}_i; \boldsymbol{\gamma}(B))\right)^2. \tag{9}$$

At this point, only $B_1(\boldsymbol{x})$ can be chosen as a parent term. Then, $B_2(\boldsymbol{x})$ and $B_3(\boldsymbol{x})$ are formed from the following expressions, respectively: $1 \cdot (x_j - t_j)_+$ and $1 \cdot (t_j - x_j)_+$. Notice that $B_1(\boldsymbol{x})$ is a 0-degree basis function, while $B_2(\boldsymbol{x})$ and $B_3(\boldsymbol{x})$ are 1-degree basis functions. From this point, $B_2(\boldsymbol{x})$ and $B_3(\boldsymbol{x})$ can already be selected as parent terms and therefore can give rise to multivariate spline basis functions. Any basis function of, at most, K_b degree with $K_b \geq 1$, is defined by the following expression:

$$B_b(\boldsymbol{x}) = \prod_{k=1}^{K_b}\left[\psi_{bk} \cdot \left(x_{j_{bk}} - t_{j_{bk}}\right)\right]_+, \forall b \in B. \tag{10}$$

Here $\psi_{bk} = \pm 1$ indicates the sense of the hinge function, $x_{j_{bk}}$ is the j-th predictor variable corresponding to the k-th term in the product for the b-th basis function, $t_{j_{bk}}$ is a value such that $t_{j_{bk}} \in \{x_{j_{bk}i} | B_b(\boldsymbol{x}_i) > 0\}$ and K_b is the number of factors that give rise to the term $B_b(\boldsymbol{x})$.

This interaction term must necessarily involve different variables to avoid producing dependencies on individual variables of a high power that can be very sensitive to extreme values. Thus, to introduce a new basis function of degree K_b, some conditions must be met: (i) a basis function of degree $K_b - 1$ must have been previously entered in the model and (ii) the same variable cannot appear twice in the product. A new basis function can be kept as univariate by selecting $B_1(\boldsymbol{x})$ as the parent term. In (10), K_b is usually limited by a hyperparameter that determines the maximum degree allowed in the interaction terms. As a general rule, it is established in 2 or 3. In case of considering $K_b = 1$, $\forall b \in B$, $B_1(\boldsymbol{x}) = 1$ would be the only possible parent term and a purely additive model with only univariate basis functions would be made. This is referred as the additive version of the MARS model. Under this scenario, by abuse of notation, we will directly write ψ_b, x_{j_b} and t_{j_b} instead of ψ_{b1}, $x_{j_{b1}}$ and $t_{j_{b1}}$.

In (5), the parameters in $\boldsymbol{\gamma}(B)$ are estimated using the least-squares method through the following Quadratic Programming model:

$$\varepsilon(\boldsymbol{\gamma}(B)) = \text{minimize}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_B(\boldsymbol{x}_i; \boldsymbol{\gamma}(B))\right)^2. \tag{11}$$

The forward algorithm creates basis functions while reducing the lack-of-fit criterion until a maximum number of terms specified by the user is reached. At the end of this first step, the estimator should overfit the data and therefore, a backward-pruning procedure is required to remove those basis functions that do not contribute significantly to the fit of the model. As Friedman (1991) describes, the regions created during the forward selection overlap and the basis function $B_1(\boldsymbol{x})$ cannot be eliminated. These two conditions prevent the discontinuity of the estimator and, in consequence, it is not necessary to use a complex pruning procedure based on sibling pairs as in CART (Breiman et al., 1984).

The backward elimination is aimed at reducing the complexity of the model built in the first step to avoid overfitting. For this

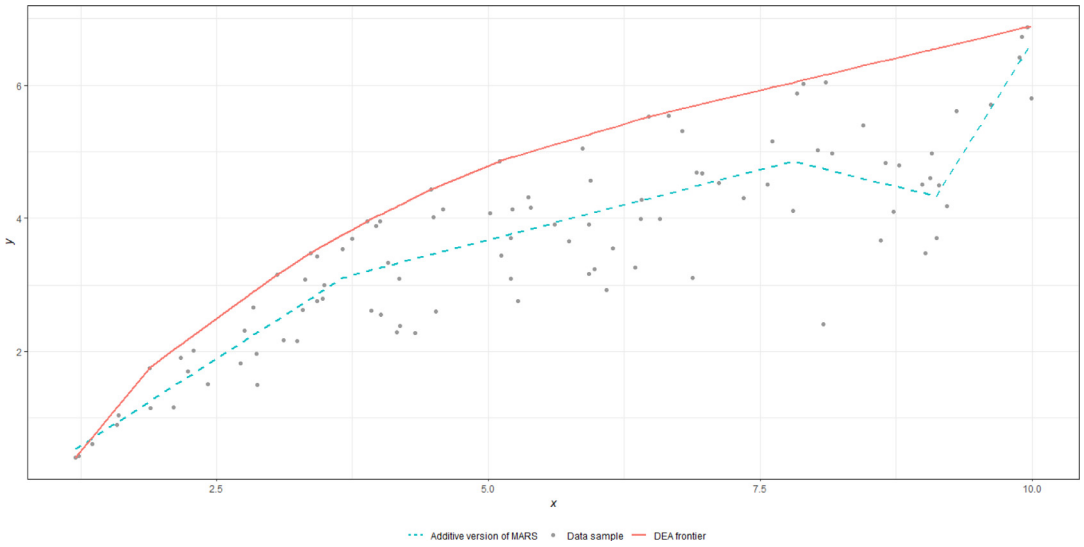

Fig. 3. An example of DEA and an additive version of MARS.

purpose, this second step prunes the model by deleting those basis functions with a lower contribution to the model's accuracy according to the generalized cross-validation (GCV) criterion (Golub et al. 1979). The GCV can be expressed as follows.

$$GCV(B) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_B(\boldsymbol{x}_i; \boldsymbol{\gamma}(B))\right]^2}{\left[1 - \frac{\tilde{C}(B)}{n}\right]^2}, \tag{12}$$

where $\tilde{C}(B)$ is a cost complexity function defined as:

$$\tilde{C}(B) = C(B) + d \cdot \chi. \tag{13}$$

Here, $C(B)$ is the number of parameters in $\hat{f}_B$, the hyperparameter d (normally set between 2 and 4) penalizes the complexity of the model and χ is the number of linearly independent basis functions in $\hat{f}_B$. Hence, the backward algorithm creates a set of $|B| - 1$ sub-models by removing basis functions one by one and selects the model, including that resulting from the forward procedure, that minimizes (12).

Finally, a graphical representation of the additive version of the MARS estimator is shown (see Fig. 3). Notice that both techniques, DEA and the additive version of MARS, share the construction of the corresponding estimator by piecewise-linear functions (as long as K_b is set to 1). Nevertheless, it is obvious that the additive MARS estimator needs certain adaptations to fulfill the usual axioms of microeconomics. A first line of action should be based on getting additive MARS to envelop the data instead of dealing with the average of the response variable. Another requirement is to get the additive MARS estimator to satisfy the properties of free disposability (monotonicity in this single-response scenario) and concavity.

3. The new approach to estimate production functions

In this section, we propose a new method based on an adaptation of the additive MARS model for estimating production functions that satisfy the usual axioms of microeconomics through a data-driven process that does not assume any particular distribution on the data noise and technical inefficiency. The new method will generate a piecewise linear function as an estimate resembling the estimator obtained through DEA. On the other hand, this new approach has an advantage over DEA in that it deals with overfitting through a pruning procedure based on generalized cross-validation (Golub et al. 1979) as in Friedman (1991).

Throughout the following sections, we review what modifications are necessary to impose on the original algorithm of the additive version of MARS to make the estimator $\hat{f}$ satisfy certain classical axioms of production theory: (A1) if $\boldsymbol{x} \leq \boldsymbol{x}'$, then $\hat{f}(\boldsymbol{x}) \leq \hat{f}(\boldsymbol{x}')$ and (A2) concavity. Postulate A1 refers to monotonicity and states that the greater amount of resources consumed by a firm, the greater the ability to produce more or at least the same output; while postulate A2 refers that $\hat{f}$ is a concave function, which is related to the convexity of the production possibility set φ in (2). Additionally, $\hat{f}$ must be a function that envelops the observations from above.

To continue, we present two subsections showing how to adapt the forward and backward algorithms associated with the additive version of MARS.

3.1. The forward algorithm

First, we introduce the two key elements that need to be adapted for the standard additive version of the MARS model to be used in the world of production function estimation:

1. Limiting the maximum degree of the basis functions (BFs) in (10) to generate a purely additive MARS model. That is, the new technique only allows univariate BFs.
2. Adding additional constraints to the programming model defined in (11) to estimate a function that envelops the data from above and satisfies both monotonicity and concavity.

We start with point 1. The satisfaction of the axioms of monotonicity (A1) and concavity (A2) can only be eased by setting a maximum degree of 1 in the construction of the set of BFs in (10). As a result, the interaction of variables (multivariate BFs) is not allowed. This limitation might compromise the predictive ability of the algorithm in some odd scenarios with continuity beyond the first derivative that cannot be fitted through 1-degree splines (Eilers and Marx, 2010); however, it provides a notable advantage in computational terms. It is easy to see that the most computationally demanding piece of code is the fitting of the parameters through the minimization problem in (11). The total computation time is proportional to the sample size, the number of predictors and the level of interactions between variables. Hence, by restricting K_b to 1, the computational cost is significantly reduced. We can name another advantage derived from this restriction. The estimator linked to this additive model is piecewise linear, thus enabling a direct comparison with the DEA estimator. In fact, somehow, the

new model could be reinterpreted as a pruned version of DEA that overcomes its overfitting problem.

Regarding point 2, the specific adaptations of the additive MARS model to satisfy axioms A1 and A2 are gradually detailed throughout the text. From Fig. 3, it can be seen that MARS was not designed by Friedman (1991) to deal specifically with production frontier estimation in microeconomics. Obviously, some well-known techniques for estimating coefficients in regression analysis, such as standard Least Squares or the Cholesky decomposition, cannot be used since they are aimed at estimating the mean value of the response variable. As an alternative to (11), we propose a linear optimization program that includes some extra constraints to capture the estimation of maximum trends instead of mean trends and that ensure that the postulates of monotonicity (A1) and concavity (A2) described above are fulfilled. In this way, a natural adaptation of the additive version of MARS algorithm to the discipline of Efficiency Analysis is achieved, although it should be noted that, the addition of new constraints to the optimization model to be solved entails a higher computational cost.

Let us recall the process of introducing a new pair of BFs in the standard MARS. The algorithm must select an input variable j, $j = 1,\dots,m$, a knot $t_j \in \{x_{j1}, x_{j2}, \dots, x_{jn}\}$ and a parent basis function with the aim of reducing the lack-of-fit criterion in the training sample. Nevertheless, the maximum degree of a BF is restricted to 1 in our approach, thus the parent term will always be $B_1(\boldsymbol{x}) = 1$. Now, for the sake of convenience, we rewrite the estimator (5) in terms of reflected pairs instead of basis functions. Accordingly, our estimator is as follows:

$$\hat{f}_P\big(\boldsymbol{x};\ \tau_0(P),\ \boldsymbol{\alpha}(P),\ \boldsymbol{\beta}(P)\big) = \tau_0(P) + \sum_{p=1}^{|P|} h_p\big(\boldsymbol{x};\ \boldsymbol{\alpha}(P),\ \boldsymbol{\beta}(P)\big)$$

$$= \tau_0(P) + \sum_{p=1}^{|P|}\Big[\alpha_p(P)\cdot\big(x_{j_p} - t_{j_p}\big)_+ + \beta_p(P)\cdot\big(t_{j_p} - x_{j_p}\big)_+\Big], \quad (14)$$

where P is the set of reflected pairs at a certain generic stage of the forward procedure (following the sequential order in which the variables were introduced in the algorithm), $\{(x_{j_p} - t_{j_p})_+, (t_{j_p} - x_{j_p})_+\}$ denotes the p-th reflected pair incorporated into the model and $\boldsymbol{\alpha}(P)$ and $\boldsymbol{\beta}(P)$ are vectors of unknown coefficients to be estimated.

Henceforth, we deal with describing the requirements necessary to comply with the conditions set out in point 2 above. The first property we satisfy refers to the enveloping nature of the production function, $\widehat{f}$, estimated through the new approach. This condition implies that, given $(\boldsymbol{x}_i, y_i)$, $\widehat{f}(\boldsymbol{x}_i)$ must necessarily be above the observed output y_i. Mathematically, this is expressed as $y_i \le \hat{f}(\boldsymbol{x}_i)$ for each learning sample $i, i = 1,\dots,n$. Therefore, it seems natural to force the estimator to meet the same association. At this point, the linear optimization program to be solved under the new approach would be as follows:

$$\underset{\boldsymbol{\varepsilon},\tau_0(P),\boldsymbol{\alpha}(P),\boldsymbol{\beta}(P)}{\text{minimize}} \quad \sum_{i=1}^{n}\varepsilon_i$$

subject to

$$\tau_0(P) + \sum_{p=1}^{|P|}\Big[\alpha_p(P)\cdot\big(x_{j_p} - t_{j_p}\big)_+ + \beta_p(P)\cdot\big(t_{j_p} - x_{j_p}\big)_+\Big]$$
$$-\varepsilon_i = y_i,\ \ i=1,\dots,n,\ \ (15.1),$$
$$\varepsilon_i \ge 0,\ \ i=1,\dots,n, \quad (15.2) \qquad (15)$$

where the new variable ε_i, which measures the error term, is defined as $\varepsilon_i = \hat{f}(\boldsymbol{x}_i) - y_i = \tau_0(P) + \sum_{p=1}^{|P|}[\alpha_p(P)\cdot(x_{j_p} - t_{j_p})_+ + \beta_p(P)\cdot(t_{j_p} - x_{j_p})_+] - y_i$. Note that the error term ε_i must be positive by constraint 15.2, and hence, there is no error compensation. Consequently, we can resort to Linear Programming rather than Quadratic Programming.

This new estimator is equivalent to the additive (forward) MARS model, but now estimating a frontier that envelops the data cloud instead of estimating the mean behavior of the data thanks to constraints (15.1) and (15.2). However, non-monotonic estimators can be given by the model (15), thereby not satisfying axiom A1. Likewise, concavity of the estimator is not guaranteed, which would be a contradiction in terms of axiom A2. We can observe these facts in Fig. 4 where an enveloping but not monotonic nor concave estimator gives rise to a non-convex technology (the shaded area).

The idea behind the satisfaction of non-decreasing monotonicity (A1) and concavity (A2) of the production function estimation with the new method is quite simple. The sum of non-decreasing monotonic functions yields a non-decreasing monotonic function and, in the same way, the sum of concave functions produces a concave function. Therefore, the strategy to follow consists of dealing with each reflected pair separately ensuring both properties, so that later, they are also satisfied by the estimator $\widehat{f}$ through the sum of non-decreasing monotonic concave functions.

Next, we establish sufficient conditions to impose monotonicity and concavity on the estimator under the new approach. In particular, as we mentioned above, we exploit the well-known result that states that the sum of several monotonically non-decreasing and concave functions is monotonically non-decreasing and concave. Consequently, the corresponding proofs of Propositions 1 and 2 are both straightforward.

Proposition 1. If $\alpha_p(P) \ge 0$ and $\beta_p(P) \le 0$, $p = 1,\dots,|P|$, then the function in (14) is monotonically non-decreasing.

Proposition 2. If $\alpha_p(P) + \beta_p(P) \le 0$, $p = 1,\dots,|P|$, then the function in (14) is concave.

Therefore, by adding these new three constraints resulting from Proposition 1 and Proposition 2 to model (15), the Linear Programming model to solve during the forward selection procedure must be the following:

$$\underset{\boldsymbol{\varepsilon},\tau_0(P),\boldsymbol{\alpha}(P),\boldsymbol{\beta}(P)}{\text{minimize}} \quad \sum_{i=1}^{n}\varepsilon_i$$

subject to

$$\begin{aligned}
&\tau_0(P) + \sum_{p=1}^{|P|}\Big[\alpha_p(P)\cdot\big(x_{j_p} - t_{j_p}\big)_+ + \beta_p(P)\cdot\big(t_{j_p} - x_{j_p}\big)_+\Big] - \varepsilon_i = y_i, && i = 1,\dots,n, && (16.1)\\
&\varepsilon_i \ge 0, && i = 1,\dots,n, && (16.2)\\
&-\alpha_p(P) - \beta_p(P) \ge 0, && p = 1,\dots,|P|, && (16.3)\\
&\alpha_p(P) \ge 0, && p = 1,\dots,|P|, && (16.4)\\
&-\beta_p(P) \ge 0, && p = 1,\dots,|P|, && (16.5)
\end{aligned} \qquad (16)$$

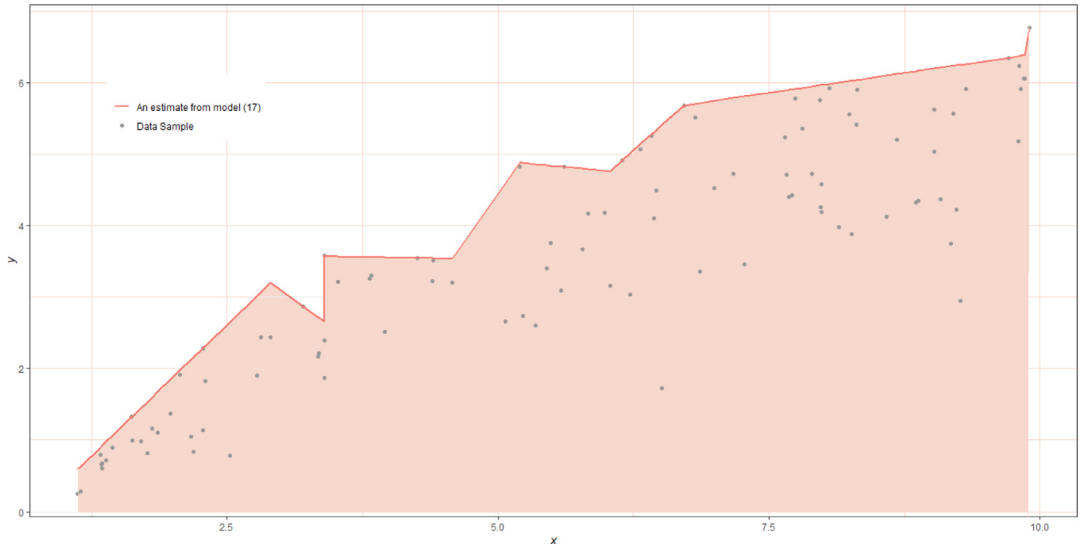

Fig. 4. A possible estimate obtained from model (15).

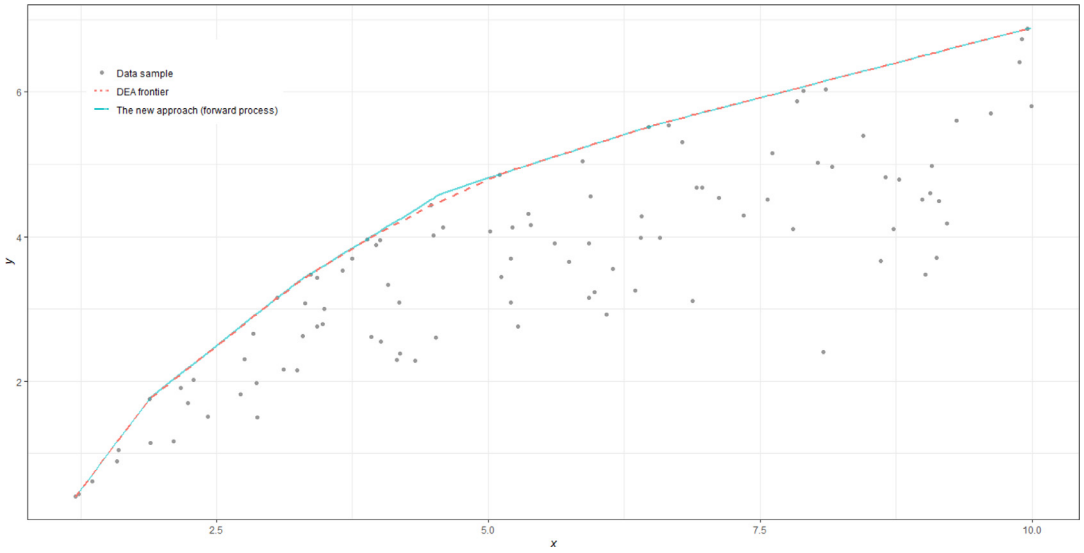

Fig. 5. An example of the frontier estimates linked to DEA and the new approach after the forward process.

The constraints (16.1) and (16.2) correspond to the restrictions of model (15). Moreover, constraints (16.3) and (16.4)-(16.5) are included to guarantee the satisfaction of concavity and monotonicity, respectively. With this, the estimator linked to model (16) fulfills the conditions set out in points 1 and 2 (at the beginning of this section). However, it still suffers from overfitting as DEA. Obviously, the accuracy at this point is fairly good since the piece-wise linear estimator closely approximates the data sample (low bias). Unfortunately, this estimator depends excessively on the training data (high variance) and therefore makes it difficult to yield a good generalization performance. These are also common features in the DEA approach. Precisely, in Fig. 5, it can be seen that DEA, which exhibits a noticeable overfitting by construction, and the new approach before pruning-back, provide almost identical estimators.

An additional difference between our method and DEA is that the new approach has a family of hyperparameters that can be tuned to obtain alternative (forward) production frontiers for the same database. Some of these parameters have already been used during the description of our algorithm. The degree of overfitting of the model can be controlled by the maximum number of pairs to be incorporated into the model (η) and the minimum reduced error rate for the addition of two new BFs (ξ). Moreover, the computational speed and the shape of the piecewise linear estimator can be regulated by (i) minspan (L), i.e., the minimum number of observations between two adjacent eligible knots, (ii) endspan (Le), i.e., the minimum number of observations before the first and after the last knot and (iii) the procedure to create the grid of eligible knots, which can be based on the observed values (as the original approach) or created ad-hoc by the user. In Fig. 6 we can see how different hyperparameterizations of the algorithm give rise to frontiers that approximate to a greater or lesser extent the training sample used. Note that, although the production frontier resulting from the new approach after executing the forward procedure will not be the final one, its shape, however, will considerably condition the final production function resulting from the backward algorithm. For example, a function too far from the data could lead to an underfitted model, while a function too close to the training sample could prevent optimal correction using the backward algorithm. In this case, Friedman (1991) recommends that back pruning discards around half of the BFs created during the first stage. As a general rule, and depending on the sample size, resampling techniques such as hold out (commonly known as training and test split) or cross-validation can be used to select the optimal set of hyperparameters.

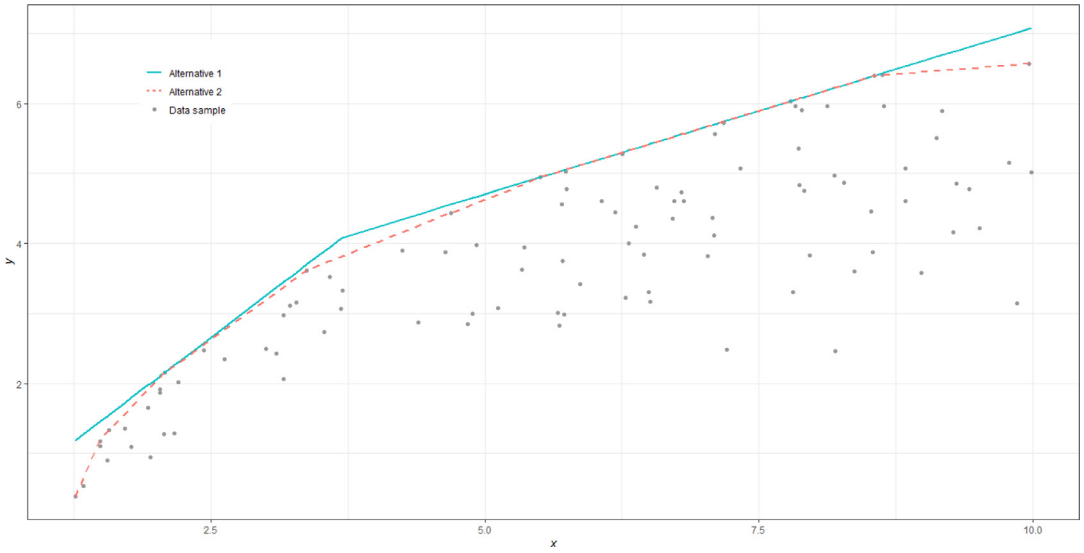

Fig. 6. An example of two alternative frontier estimates linked to the new approach after the forward process.

Algorithm 1
Forward procedure for the new approach.

Algorithm 1. Forward procedure for the new approach

Input:

η, maximum number of reflected pairs

ξ, minimum error reduction for a new iteration to be performed

Output: $f_{\tilde{P}^*}$; P^*

$LOF_{\min} := \infty$; $\tilde{P} := \varnothing$
WHILE $|\tilde{P}| < \eta$
  $LOF'_{\min} := LOF_{\min}$
  FOR $j = 1$ **TO** m
    FOR $i = 1$ **TO** n
      Select a knot: $t_j := x_{ji}$
      Consider a new reflected pair into model (14): $\left\{(x_j - t_j)_+, (t_j - x_j)_+\right\}$
      Update $\tilde{P}$ as $\tilde{P} \cup \left\{\left(x_{j|\tilde{P}|+1} - t_{|\tilde{P}|+1}\right)_+, \left(t_{|\tilde{P}|+1} - x_{|\tilde{P}|+1}\right)_+\right\}$ with $x_{j|\tilde{P}|+1} := x_j$ and $t_{j|\tilde{P}|+1} := t_j$
      Obtain the coefficients $\tau_0(\tilde{P})$, $\boldsymbol{\alpha}(\tilde{P})$, $\boldsymbol{\beta}(\tilde{P})$ by (16).
      Determine LOF by (9).
      IF $LOF < LOF'_{\min}$ **THEN**
        $LOF'_{\min} := LOF$;
        $P_{opt} := \tilde{P}$
      END IF
    END FOR
  END FOR
  $\tilde{P} := P_{opt}$
  IF $LOF'_{\min} < LOF_{\min} \cdot (1 - \xi)$ **THEN**
    $LOF_{\min} := LOF'_{\min}$
  ELSE
    end loop
  END IF
END WHILE
$P^* := \tilde{P}$
$f_{\tilde{P}^*}$ is defined as in (14) with $P = P^*$

Finally, the steps that must be carried out in the forward procedure to determine a frontier estimate linked to the new approach are shown in Algorithm 1, where P^* represents the set of reflected pairs at the end of the forward procedure.

3.2. The backward algorithm

Overfitting is a key threat to the reliability of a statistical model. The new technique, as in the original MARS algorithm, makes intensive use of the response variable to define the set of BFs. This fact, in general, drastically reduces the bias of the model, but at the same time increases its variance. It means that the model may "memorize" the training data and, in consequence, not be able to provide a good response to a new sample. This is a common problem in machine learning algorithms. Conveniently, these algorithms always include certain procedures to accomplish the model to generalize correctly. For our purpose, we suggest applying our approach along with a pruning procedure to suitably evaluate production frontiers. To do so, the standard approach performed by Friedman (1991) in MARS based on generalized cross-validation will be slightly adapted to meet the requirements of the frontier analysis framework.

The forward stepwise procedure ends with the creation of a set of paired BFs, in addition to the constant basis function $B_1(\boldsymbol{x}) = 1$ (intercept term). This model generally suffers from overfitting (see Fig. 5). Consequently, a process of backward elimination is initiated (see Section 2.2) where those BFs that do not contribute significantly to the improvement of the model's performance are discarded. In other words, it attempts to promote an optimal balance between the complexity and the precision of the model. Naturally, this approach breaks the reflected pair structure used during the first stage of the algorithm. Now, only some pairs of BFs will be kept in the model, while others will be totally or partially eliminated. With this, each BF can be in three different states: paired, left-side unpaired and right-side unpaired. In this way, function (5) can be redefined as follows:

$$\begin{aligned}\widehat{f}_B\left(\boldsymbol{x};\ \tau_0(B),\ \boldsymbol{\alpha}\ (B),\ \boldsymbol{\beta}(B),\ \boldsymbol{\delta}(B),\ \boldsymbol{\omega}(B)\right) &= \tau_0(B) + \sum_{a=1}^{|H|} h_a\left(\boldsymbol{x};\ \boldsymbol{\alpha}\ s(B), \boldsymbol{\beta}(B)\right) + \sum_{c=1}^{|G|} g_c\left(\boldsymbol{x};\ \boldsymbol{\delta}(B)\right) \\ &\quad + \sum_{e=1}^{|R|} r_e(\boldsymbol{x};\ \boldsymbol{\omega}\ (B)),\end{aligned} \tag{17}$$

where $B = \{B_1(\boldsymbol{x})\} \cup H \cup G \cup R$ is the set of BFs that the model contains, being H the set of reflected pairs, G the set of right-side unpaired BFs and R the set of left-side unpaired BFs. In this way, $h_a(\boldsymbol{x}; \boldsymbol{\alpha}(B), \boldsymbol{\beta}(B)) = \alpha_a(B)(x_{j_a} - t_{j_a})_+ + \beta_a(B)(t_{j_a} - x_{j_a})_+$ is the a-th reflected pair in H, $g_c(\boldsymbol{x}; \boldsymbol{\delta}(B)) = \delta_c(B)(x_{j_c} - t_{j_c})_+$ is the c-th right-side unpaired basis function in G and $r_e(\boldsymbol{x}; \boldsymbol{\omega}(B)) = \omega_e(B)(t_{j_e} - x_{j_e})_+$ is the e-th left-side unpaired basis function in R.

Next, we state the conditions necessary to guarantee that the function (17) satisfies monotonicity and concavity. The proof is based on the well-known result that establishes that the sum of

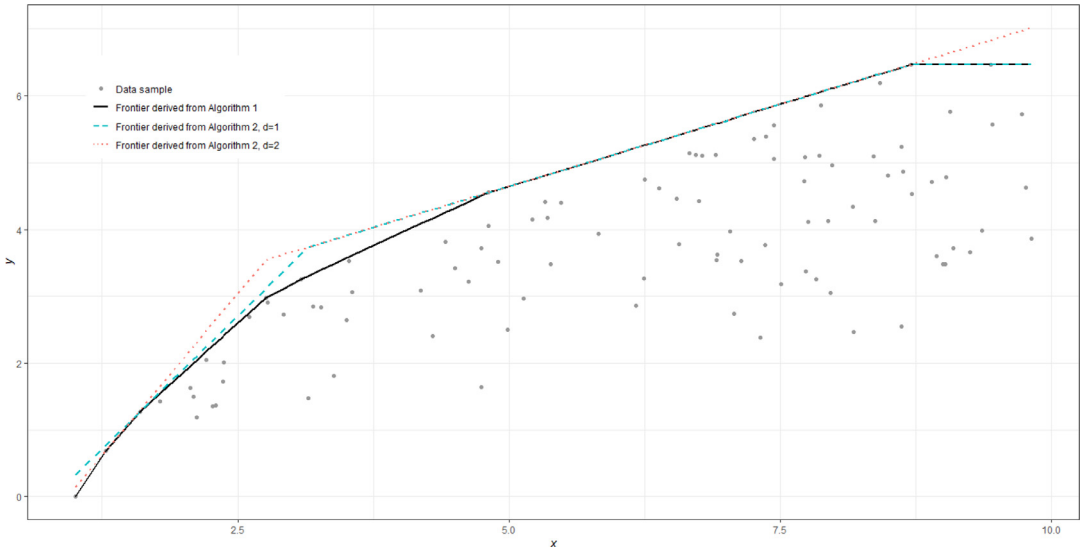

Fig. 7. An example of alternative production frontiers derived from the new approach.

non-decreasing and concave functions is also a non-decreasing and concave function.

Proposition 3. If $\alpha_a(B)+\beta_a(B)\leq 0,\ \alpha_a(B)\geq 0,\ \beta_a(B)\leq 0,\ a=1,\ldots,|H|,\ \delta_c(B)=0,\ c=1,\ldots,|G|$, and $\omega_e(B)\leq 0,\ e=1,\ldots,|R|$, then the function (17) is a non-decreasing and concave function.

Under the assumptions of the above result, notice that if $\delta_c(B)=0,\ c=1,\ldots,|G|$, then $g_c(\boldsymbol{x};\boldsymbol{\delta}(B))=0,\ c=1,\ldots,|G|$. Consequently, $g_c(\boldsymbol{x};\boldsymbol{\delta}(B)),\ c=1,\ldots,|G|$, disappears from expression (17).

From Proposition 3, we confirm the need to modify the way of proceeding described in Friedman (1991). The backward algorithm will be identical to that described in Section 2.2. except for the manner of selecting a BF to be removed. While in the standard MARS any BF is a candidate to be eliminated, in the case of the new approach, we must introduce two conditions that must be considered before selecting a BF:

1- Right-side BFs can only be removed from reflected pairs.
2- Left-side BFs can only be removed when appearing unpaired.

In this way, it is ensured that $g_c(\boldsymbol{x};\boldsymbol{\delta}(B))=0,\ \forall c=1,\ldots,|G|$.

Now, we can establish the Linear Programming model to solve during the backward stage.

$$\underset{\boldsymbol{\varepsilon},\tau_0(B),\boldsymbol{\alpha}(B),\boldsymbol{\beta}(B),\boldsymbol{\omega}(B)}{\text{minimize}} \quad \sum_{i=1}^{n}\varepsilon_i$$

subject to

$$\begin{aligned}
\tau_0(B)+\sum_{a=1}^{|H|}\left[\alpha_a(B)\left(x_{j_a i}-t_{j_a}\right)_+ +\beta_a(B)\left(t_{j_a}-x_{j_a i}\right)_+\right]+\sum_{e=1}^{|R|}\left[\omega_e(B)\left(t_{j_e}-x_{j_e i}\right)_+\right]-\varepsilon_i &= y_i, && i=1,\ldots,n, && (18.1)\\
\varepsilon_i &\geq 0, && i=1,\ldots,n, && (18.2)\\
-\alpha_a(B)-\beta_a(B) &\geq 0, && a=1,\ldots,|H| && (18.3)\\
\alpha_a(B) &\geq 0, && a=1,\ldots,|H| && (18.4)\\
-\beta_a(B) &\geq 0, && a=1,\ldots,|H| && (18.5)\\
-\omega_e(B) &\geq 0, && e=1,\ldots,|R| && (18.6)
\end{aligned} \qquad (18)$$

Model (18) only includes a new constraint, (18.6), with respect to the forward model (16), which makes it possible to ensure that the left-side unpaired BFs also comply with the monotonicity and concavity properties.

The steps that must be carried out in the backward procedure of the new approach are shown in Algorithm 2.

Finally, Fig. 7 shows the effect of the pruning procedure on the function obtained after the forward algorithm. For this, two different values of the hyperparameter d have been used ($d=1$ and $d=2$). The quantity d in (13) represents a cost for each BF that is maintained in the model. Larger values for d lead to a smaller number of knots being placed and thereby a model less prone to suffer from overfitting. Again, the optimal value of d can be optimally selected by hold-out or cross-validation.

4. Computational experience

Here we describe the simulation results that allow the comparison of the following methods: the frontier estimate derived from Algorithm 2, DEA and C^{2}NLS. Thus, an assessment of these techniques carried out under six different simulated scenarios is presented. These same six scenarios were defined by Kuosmanen and Johnson (2010). Their descriptions appear in Table 1.

Scenarios 1 and 2 represent a single-input case, while scenarios 3-6 represent multi-input cases with interaction of variables (two and three inputs with different curvatures for the target function). For all scenarios, we tested three data set sizes of 50, 100 and 150 observations. The input data were randomly sampled from a uniform distribution $Uni[1,10]$, independently for each input and firm. Subsequently, a random inefficiency term $u\sim|N(0,0.4)|$ was com-

Table 1
Simulated scenarios.

Scenario	# Inputs	Target function $f(\boldsymbol{x})$
1	1	$\ln(x_1)+3$
2	1	$3+x_1^{0.5}+\ln(x_1)$
3	2	$0.1x_1+0.1x_2+0.3(x_1x_2)^{0.5}$
4	3	$0.1x_1+0.1x_2+0.1x_3+0.3(x_1x_2x_3)^{1/3}$
5	2	$0.1x_1+0.1x_2+0.3(x_1x_2)^{1/3}$
6	3	$0.1x_1+0.1x_2+0.1x_3+0.3(x_1x_2x_3)^{1/4}$

Algorithm 2
Backward procedure for the new approach.

Algorithm 2. Backward procedure for the new approach

Input:

f_{P^*}, the model after the forward procedure.

P^*, set of reflected pairs after the forward procedure.

Output: $f_{B^*}; B^*$

The set of paired BFs in P^*: H
The set of right-side paired BFs in H : H_R
The set of left-side unpaired BFs $R \coloneqq \varnothing$
$\tilde{\boldsymbol{B}} \coloneqq \{\boldsymbol{B}_1(\boldsymbol{x})\} \cup H \cup R$
Obtain the coefficients $\tau_0(\tilde{\boldsymbol{B}})$, $\boldsymbol{\alpha}(\tilde{\boldsymbol{B}})$, $\boldsymbol{\beta}(\tilde{\boldsymbol{B}})$, $\boldsymbol{\omega}(\tilde{\boldsymbol{B}})$ by (18).
Compute $GCV(\tilde{\boldsymbol{B}})$ by (12).
WHILE $|\tilde{\boldsymbol{B}}| > 1$
  $GCV_{\min} \coloneqq \infty$
  Define the set of BFs to be selected for being removed: $W \coloneqq H_R \cup R$
  FOR EVERY $w \in W$
    IF $w \in H_R$ **THEN**
      $H'_R \coloneqq H_R \setminus \{w\}$
      Let w^* be the left-side paired BF associated with w
      $H' \coloneqq H \setminus \{w, w^*\}$
      $R' \coloneqq R \cup \{w^*\}$
    ELSE
      $R' \coloneqq R \setminus \{w\}$
    END IF
    Update $\tilde{\boldsymbol{B}} \coloneqq \{\boldsymbol{B}_1(\boldsymbol{x})\} \cup H' \cup R'$
    Compute $GCV(\tilde{\boldsymbol{B}})$ by (12).
    IF $GCV(\tilde{\boldsymbol{B}}) < \boldsymbol{GCV}_{\min}$ **THEN**
      $GCV_{\min} \coloneqq GCV(\tilde{\boldsymbol{B}})$
      $B_{opt} = \tilde{\boldsymbol{B}}$, $H'_{opt} = H'$, $H'_{R_{opt}} = H'_R$, $R'_{opt} = R'$
    END IF
  END FOR
  $\tilde{\boldsymbol{B}} = \boldsymbol{B}_{opt}$, $H = H'_{opt}$, $H_R = H'_{R_{opt}}$, $R = R'_{opt}$
END WHILE
$B^* \coloneqq \tilde{\boldsymbol{B}}$ with $GCV(\tilde{\boldsymbol{B}}) = \boldsymbol{GCV}_{\min}$
f_{B^*} is defined as in (17) with $B = B^*$ and $\sum_{c=1}^{|G|} g_c(x; \boldsymbol{\delta}(B^*)) \coloneqq 0$

puted. Then, the output used for the analysis was calculated as $y = f(\boldsymbol{x}) - u$. We ran 100 trials ($l = 1, \ldots, 100$) for each combination of scenario and data set size to investigate the relative performance of the methods. Performance of each method was evaluated by two standard criteria: the mean squared error (MSE) and the bias. The MSE statistic is defined as $\sum_{l=1}^{100}\sum_{i=1}^{n}(\widehat{f}(\boldsymbol{x}_i^l) - f(\boldsymbol{x}_i^l))^2/100n$, while the bias is computed as $\sum_{l=1}^{100}\sum_{i=1}^{n}(\widehat{f}(\boldsymbol{x}_i^l) - f(\boldsymbol{x}_i^l))/100n$, where $\boldsymbol{x}_i^l$ denotes the i-th input profile corresponding to the l-th trial. At this point, let us highlight two details. First, in these two formulas, $f(\boldsymbol{x}_i^l)$ denotes the value of the true frontier while $\widehat{f}(\boldsymbol{x}_i^l)$ represents its estimation for the input profile $\boldsymbol{x}_i^l$. Second, in the formula corresponding to the bias, it is more usual to resort to the absolute value of the differences. However, at this point, we follow Kuosmanen and Johnson (2010), from which we mimic the simulation scenarios with the objective of comparing the results, where the authors defined the bias in this way to identify the 'sign' of the deviation: negative ($\widehat{f} < f$) or positive ($\widehat{f} > f$). In particular, the accuracy of the estimates in quadratic terms is measured by the MSE, allocating the same weight to negative and positive deviations. Therefore, MSE will be used as a model evaluation metric. The bias statistic, instead, indicates whether the estimated frontier $\widehat{f}$ systematically underestimates ($bias < 0$) or overestimates ($bias > 0$) the true frontier f. We note that positive and negative deviation terms cancel out when averaged over the observations and simulation runs; however, it does give useful information about the behavior of the estimated frontier with respect to the target frontier. Then, following Kuosmanen and Johnson (2010), we analyze the model's performance in two ways: the magnitude (MSE) and the direction (bias) of the error.

Additionally, we determined the best set of hyperparameters for each trial by a training (70%) and test (30%) split due to the high computational cost involved in cross-validation. In our context, the hyperparameters are η, ξ, L, Le and d. The grid of available knots was the observed data as in Friedman (1991). We fixed the maximum number of pairs to be incorporated into the model (η) and thereby we only control the growth of the forward algorithm through ξ. From our own experience, we set the following finite value space for each hyperparameter: $\xi \in \{0.1, 0.01, 0.001, 0\}$, $L \in \{-2, -1\}$, $Le \in \{-2, -1\}$ and $d \in \{1, 2, 3\}$. The -1 and -2 values in L and Le correspond to the minspan and endspan approaches in Friedman (1991) and Zhang (1994), respectively. These values generated a total of 48 different hyperparameter combinations for the proposed simulations.

Table 2 records the mean, the standard deviation (in brackets) and the median of the best-performing hyperparameters in our simulations. The results are detailed below. The best value of the ξ hyperparameter is highly dependent on the sample size. Specifically, it reveals an inverse relationship between the number of observations in the sample and the optimal value of ξ. The value 0.1 does not seem to allow the (forward) model to grow enough and therefore does not provide promising results. In this case, it seems reasonable to search for optimal values near 0.01 (above and below) since values around 0.001 represent an increase in computational cost that does not necessarily improve the results obtained. Regarding the hyperparameters L and Le, a certain general tendency is observed for the value -1 (Friedman approach) in case of the minspan. The Friedman and Zhang approaches provide very similar results in case of the endspan, therefore, they can be chosen interchangeably. Regarding the hyperparameter d, the value 3 could be discarded since they do not usually perform well. Likewise, value 1 seems to give the best results.

Table 2 also shows the computation time associated with the new technique. It should be noted that the computing time spent can be seen as a drawback of the new approach in comparison with other techniques such as DEA, which is directly based on Linear Programming. The experiments were conducted on a workstation with 2.3 GHz Intel(R) Xeon(R) CPU E5-2650 v3 with 40 cores, 62 Gigabyte of RAM and an Ubuntu18.04.5 LTS operating system. The code was implemented in R version 3.6.3. The code is hosted in an open-source repository on GitHub at https://github.com/Victor-Espana/MLFrontiers. To solve the optimization problems (16) and (18), the Rglpk package (Theussl and Hornik, 2019) was utilized. Concerning the execution time, the results illustrate an exponential relationship with the sample size. Moreover, this situation becomes even more critical when the number of inputs is also increased[3].

Table 3 examines the performance of the new technique, DEA and C^{2}NLS grounded on the MSE criterion. The first two columns indicate the background and the number of observations. The next four columns state the MSE means for the methods under consideration. Finally, the last four columns show the variation in the MSE statistic (and its sign) between the new technique versus DEA and C^{2}NLS. The results obtained show, in general, how the approach we have proposed present significant improvement over the DEA and the C^{2}NLS techniques. First of all, it should be men-

[3] We also tried to find out the execution time required to get the results by applying the new method when a standard desktop PC (Personal Computer) is used in the case of analyzing the most complex scenario (four variables and 150 observations). In this case, the experiments were conducted on a PC with 1.80 GHz Intel (R) Core (TM) i7-8550U CPU with 12 Gigabyte of RAM and a Windows 11 Home 64 bits operating system. We executed 100 trials, obtaining that the mean time required by our technique was 62 seconds (approximately one minute).

Table 2
Optimal hyperparameters for the new approach.

Scenario (#inputs)	Sample size	Hyperparameter								Execution time (in seconds)
		ξ		L		Le		d		
		Mean (std)	Median	Mean (std)	Median	Mean (std)	Median	Mean (std)	Median	
1 (1)	50	0.059 (0.048)	0.100	-1.33 (0.47)	-1	-1.47 (0.50)	-1	1.45 (0.64)	1	0.32
	100	0.025 (0.040)	0.001	-1.36 (0.48)	-1	-1.35 (0.48)	-1	1.42 (0.70)	1	2.73
	150	0.011 (0.027)	0.001	-1.40 (0.49)	-1	-1.25 (0.44)	-1	1.24 (0.55)	1	6.93
2 (1)	50	0.043 (0.048)	0.010	-1.29 (0.46)	-1	-1.40 (0.49)	-1	1.48 (0.75)	1	0.46
	100	0.014 (0.031)	0.001	-1.45 (0.50)	-1	-1.37 (0.49)	-1	1.24 (0.53)	1	3.44
	150	0.004 (0.011)	0.001	-1.42 (0.50)	-1	-1.38 (0.49)	-1	1.22 (0.46)	1	9.16
3 (2)	50	0.045 (0.047)	0.010	-1.37 (0.49)	-1	-1.61 (0.49)	-2	1.84 (0.85)	2	1.05
	100	0.038 (0.046)	0.010	-1.36 (0.48)	-1	-1.48 (0.50)	-1	1.64 (0.84)	1	4.69
	150	0.028 (0.041)	0.010	-1.21 (0.41)	-1	-1.32 (0.47)	-1	1.65 (0.82)	1	11.02
4 (3)	50	0.034 (0.045)	0.010	-1.42 (0.50)	-1	-1.52 (0.50)	-2	1.92 (0.87)	2	2.16
	100	0.018 (0.034)	0.001	-1.34 (0.48)	-1	-1.51 (0.50)	-2	1.42 (0.68)	1	14.11
	150	0.016 (0.030)	0.010	-1.32 (0.47)	-1	-1.55 (0.50)	-2	1.47 (0.73)	1	36.72
5 (2)	50	0.046 (0.047)	0.010	-1.31 (0.46)	-1	-1.51 (0.50)	-2	1.97 (0.87)	2	1.00
	100	0.042 (0.047)	0.010	-1.24 (0.43)	-1	-1.34 (0.48)	-1	1.56 (0.77)	1	4.73
	150	0.034 (0.044)	0.010	-1.27 (0.45)	-1	-1.33 (0.47)	-1	1.50 (0.70)	1	12.12
6 (3)	50	0.036 (0.046)	0.010	-1.27 (0.45)	-1	-1.60 (0.49)	-2	2.07 (0.77)	2	2.32
	100	0.024 (0.039)	0.010	-1.40 (0.49)	-1	-1.39 (0.49)	-1	1.56 (0.74)	1	12.50
	150	0.020 (0.037)	0.001	-1.40 (0.49)	-1	-1.36 (0.48)	-1	1.43 (0.69)	1	39.46

Table 3
Relative performance of estimation methods linked to MSE.

Scenario (#inputs)	Sample size	Mean squared error			Variation in MSE (%)	
		The new approach	DEA	C^2NLS	The new approach vs DEA	The new approach vs C^2NLS
1 (1)	50	0.007	0.010	0.006	-30.45	+14.38
	100	0.003	0.005	0.005	-44.59	-46.78
	150	0.002	0.003	0.004	-35.71	-55.33
2 (1)	50	0.007	0.011	0.009	-32.99	-15.64
	100	0.004	0.006	0.006	-41.08	-36.82
	150	0.002	0.003	0.006	-36.81	-63.08
3 (2)	50	0.018	0.030	0.013	-39.29	+38.52
	100	0.018	0.018	0.009	-03.39	+86.79
	150	0.018	0.012	0.010	+51.86	+76.32
4 (3)	50	0.018	0.065	0.024	-71.86	-23.77
	100	0.016	0.046	0.017	-65.11	-06.80
	150	0.015	0.035	0.015	-56.08	+01.12
5 (2)	50	0.008	0.029	0.014	-72.39	-41.79
	100	0.004	0.015	0.008	-73.57	-53.96
	150	0.003	0.011	0.007	-75.04	-60.51
6 (3)	50	0.012	0.060	0.022	-80.30	-46.59
	100	0.006	0.040	0.016	-84.84	-60.52
	150	0.004	0.030	0.014	-86.23	-69.50

tioned that all the techniques are affected by the increase in dimensionality, since the mean MSE increases when the number of inputs increases. However, this does not occur in the same proportion for all methods. While DEA increases the MSE by 646% by increasing the number of inputs from 1 to 3 (scenarios 1 and 2 versus scenarios 4 and 6), in the case of our approach, this increase is between 3 and 4 times lower. Therefore, it seems that the new technique is more robust than DEA to the curse of dimensionality. The improvements of the new approach over DEA ranged from 3.39% to 86.23%. Scenario 3 is the most unfavorable for our approach, especially in the case of 150 samples. It is worth noting that the additive nature of our model can be a limitation depending on the degree of curvature of the true production function considered. The production functions associated with scenarios 3 and 5 are the same except for the value of the exponent corresponding to the interaction term between input 1 and input 2. While our results are poor in scenario 3, they are really good in the case of scenario 5. The reason is the value of the exponent. We executed an extra computational experience following the same mathematical expression for the production function as that used in scenarios 3 and 5 but playing with different values for the exponent, between 0.1 and 1. Our results showed that the new approach seems to work well up to a certain value of the exponent (a threshold of around 0.5) for which the performance drops off sharply regardless of the sample size. Anyway, although our model is additive in nature, which represents a weakness from a methodological point of view when the true production function has interactions between variables, we have shown using the simulation scenarios taken from Kuosmanen and Johnson (2010) (which contains 4 non-additive scenarios with different degrees of curvature) that the new method can outperform the other two techniques considered (DEA and C^2NLS) even in non-additive situations. Indeed, considerable improvements are observed in the rest of the scenarios analysed, especially in 4, 5 and 6. Regarding C^2NLS, the improvement in results is also quite substantial. In this case, the improvement percentages of the new technique with respect to C^2NLS is between 6.80% and 69.50%. The conclusions reached are very similar to those described above. C^2NLS provides better results for all sample sizes in scenario 3, while in the rest of the cases, our technique performs better. In the 1-input scenarios (1 and 2), a greater

Table 4
Relative performance of estimation methods linked to bias.

Scenario (#inputs)	Sample size	Bias		
		The new approach	DEA	C^2NLS
1 (1)	50	-0.036	-0.070	+0.006
	100	-0.021	-0.044	+0.005
	150	-0.017	-0.035	+0.004
2 (1)	50	-0.031	-0.074	+0.009
	100	-0.024	-0.051	+0.006
	150	-0.018	-0.037	+0.006
3 (2)	50	+0.002	-0.124	+0.013
	100	+0.034	-0.092	+0.009
	150	+0.049	-0.072	+0.010
4 (3)	50	-0.019	-0.199	+0.024
	100	+0.010	-0.162	+0.017
	150	+0.032	-0.140	+0.015
5 (2)	50	-0.042	-0.120	+0.014
	100	-0.021	-0.081	+0.008
	150	-0.012	-0.067	+0.007
6 (3)	50	-0.055	-0.190	+0.022
	100	-0.029	-0.149	+0.016
	150	-0.012	-0.124	+0.014

percentage of improvement is observed when the number of units in the sample increases. However, this situation is completely the opposite in scenario 4.

Table 4 reports the results based on bias. The structure of Table 4 is the same as Table 3 but does not include the last 2 columns. We first start by describing DEA's performance. The production function estimated by DEA is consistently below the true production function as reflected by all negative values in its bias. In this way, DEA is able to provide a correct description of the data but fails to provide an adequate estimate of the underlying function. That is, DEA can be considered a purely descriptive technique with little inferential power, except in the case of invoking the consistency property of the DEA estimator. Furthermore, as was the case with the MSE statistic, a decreasing trend is also observed (approaching zero) as the number of observations in the sample increases. As for the rest of the techniques, certain patterns are shown in the results obtained. The new technique tends to underestimate the true production function and also provide better results (in terms of bias) when the sample size increases. Nevertheless, the new approach overestimates the true production function in scenario 3. However, since the new technique does aim to estimate the underlying function, the presence of both positive and negative deviations is natural. On the other hand, C^2NLS tends to overestimate the true production function, and in this case, the level of bias seems to increase as the sample size grows in some scenarios such as 3 and 4.

Next, we provide the contour plot for different levels corresponding to several functions: the Cobb-Douglas production function $f(x_1, x_2) = x_1^{0.4} \cdot x_2^{0.1}$, the DEA estimate and the estimate based on the new approach (see Fig. 8). Fig. 8 illustrates the fact that DEA usually fails to adequately approach the theoretical production function, i.e., the Cobb-Douglas function in this case. In contrast, the fig. also shows that the new method is capable of achieving better approximations.

The statement that DEA (and related non-parametric envelopment techniques such as Free Disposal Hull, FDH) suffers from an overfitting problem is very recent in the literature. From a statistical point of view, overfitting is a problem that happens when you have a perfect fit of your model on the data sample. When this occurs, the model unfortunately cannot perform accurately against unseen data, which is usually related to a high generalization error (Hastie et al., 2009). Standard machine learning techniques aim to identify the actual function that lies behind the Data Generating Process (see for example Vapnik, 1998). If the precise equilibrium is struck between the ability of the model to learn any dataset without error and the accuracy achieved on a particular dataset (the observations), then an appropriate estimation of the underlying function to be estimated will be attained. This ability to learn any possible dataset is linked to the notion of the generalization error (also called out-of-sample error in the literature). In the non-parametric framework, the theoretical generalization error of a model cannot be calculated in general, but it may be approximated by resorting to test samples or cross-validation. In particular, in the context of efficiency measurement, envelopment techniques as DEA, which place the efficient frontier as close as possible to the data sample due to the minimal extrapolation principle, can correctly measure efficiency for a particular set of observations (DMUs) following a sample-specific-based evaluation, but, at the same time, suffer from overfitting. The DEA model is too close to the DMUs when the (underlying) efficient frontier is actually located above the data cloud. This last feature limits its inferential capability, at least for small data samples, a point that is important when one of the objectives of the study is stating something about the underlying function behind the Data Generating Process that produced the observations. One direct impact of this overfitting problem on the results determined through DEA is that an important part of the DMUs under evaluation are shown as technically efficient (i.e., the corresponding technical score equals one, when in fact they are not located on the underlying production frontier) and, in general, DEA scores are overly optimistic. Some very recent approaches have highlighted the overfitting problem suffered by FDH and DEA (see Esteve et al., 2020, Tsionas, 2022, Valero-Carreras et al., 2022, and Molinos-Senante et al., 2023). In this section, we have showed how DEA and the new approach estimate the target function $f(\boldsymbol{x})$ through $\hat{f}(\boldsymbol{x})$, given an input profile $\boldsymbol{x}$. Next, and as a complement, we are going to show how close the estimates $\hat{f}(\boldsymbol{x}_i)$ are to the observed output level for each DMU_i, i.e., y_i, $i = 1, \ldots, n$. Our objective is to illustrate how DEA overfits the data in comparison with the new approach. In the considered single-output production setting, a natural (output-oriented) technical efficiency score can be defined as the ratio $\phi_i = \hat{f}(\boldsymbol{x}_i)/y_i$, which compares these two quantities, i.e., the estimate $\hat{f}(\boldsymbol{x}_i)$ and the observation y_i. For this reason, and with the objective of illustrating the overfitting problem suffered by DEA in comparison with the new approach, we calculated the mean and variance of the efficiency score (ϕ_i) determined by both Data Envelopment Analysis and the additive model based on regression splines (see Table 5). The results demonstrate that the DEA estimates are closer to the

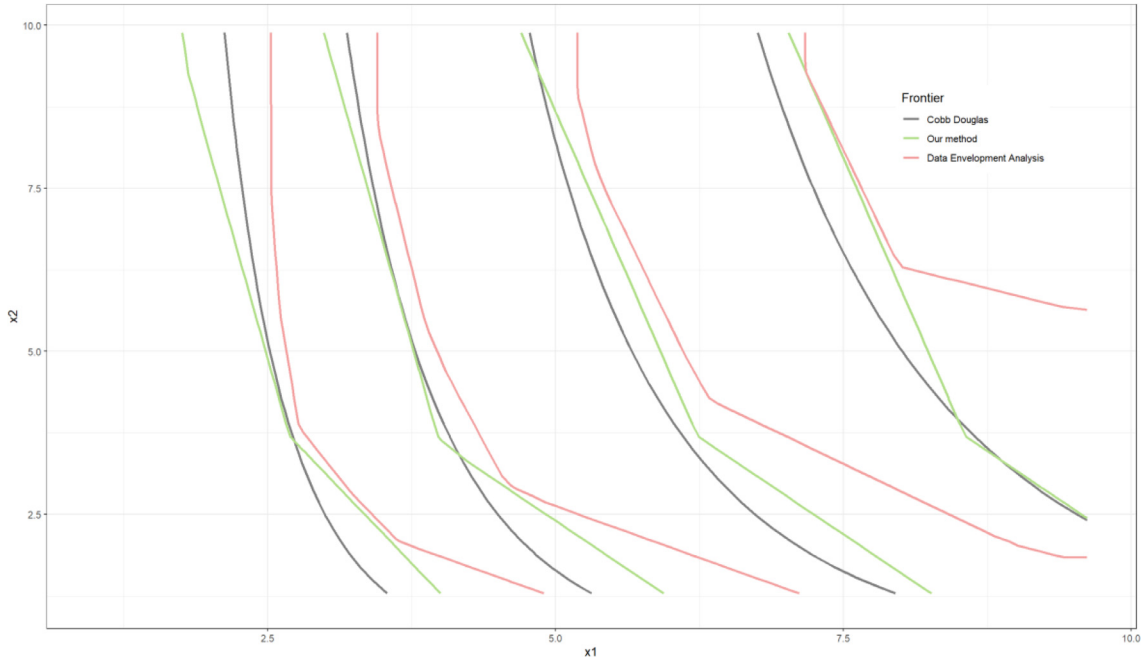

Fig. 8. Contour plot comparing the performance of the new approach with DEA in an example with two inputs and one output.

Table 5
Comparison of efficiency scores between the new approach and DEA.

Scenario (#inputs)	Sample size	The new approach		DEA		Mean difference
		Mean	Std. dev.	Mean	Std. dev.	
1	50	1.069	0.010	1.064	0.013	0.54%
	100	1.072	0.007	1.068	0.008	0.37%
	150	1.074	0.006	1.072	0.007	0.28%
2	50	1.043	0.006	1.039	0.008	0.41%
	100	1.047	0.005	1.044	0.006	0.29%
	150	1.048	0.004	1.046	0.004	0.23%
3	50	1.211	0.105	1.122	0.103	8.90%
	100	1.227	0.061	1.146	0.064	8.16%
	150	1.222	0.036	1.150	0.036	7.15%
4	50	1.123	0.033	1.047	0.033	7.57%
	100	1.136	0.020	1.064	0.023	7.18%
	150	1.150	0.037	1.079	0.039	7.09%
5	50	1.272	0.120	1.194	0.123	7.82%
	100	1.336	0.258	1.245	0.279	9.08%
	150	1.315	0.284	1.249	0.289	6.62%
6	50	1.140	0.029	1.066	0.031	7.42%
	100	1.156	0.023	1.090	0.024	6.59%
	150	1.164	0.021	1.102	0.023	6.27%

observed outputs than those provided by the new approach (with efficiency scores closer to one on average in the case of DEA). However, it should be noted that the differences in terms of overfitting between the new approach and DEA in the contexts of one and two inputs are somewhat minor. By contrast, when we increase the number of inputs being evaluated, we observe differences by up to 9% on average. For example, with 5 inputs and 100 units, DEA suggests that the output produced could be increased by 25%, while the new approach suggests that the level of output should augment by 34% (on average). Note that we have already demonstrated that the new method yields estimates closer to the underlying production function than DEA (with smaller bias and mean squared error in general). In this way, it seems that, in the simulations, DEA presents a better fit of the frontier model on the data sample in comparison with the additive model based on regression splines but at the expense of having worse estimates of the underlying Data Generating Process (the actual production function).

5. Conclusions and lines for future work

This paper has built a new bridge between production theory and machine learning in the literature. So far, these two fields have been growing exponentially but side by side and separately. However, the new tendency shows a prospective integration of the efficiency analysis world into the context of machine learning (see, for example, the recent papers by Esteve et al., 2023, Esteve et al., 2020, Valero-Carreras et al., 2021, or Olesen and Ruggiero, 2022). In our case, this has allowed us to introduce a new technique for estimating production functions through splines and recursive partitioning. The new technique endows an additive version of the Multivariate Adaptive Regression Splines (MARS) technique by Friedman (1991) with shape constraints to estimate a surface that envelops all the data from above and satisfies monotonicity and concavity. These features are linked to the traditional nonparametric Data Envelopment Analysis (DEA).

DEA and the new approach share some characteristics, such as their non-parametric nature and the piece-wise linear shape of their estimators. Nevertheless, while DEA suffers from overfitting by construction due to the axiom of minimal extrapolation assumed in the classical literature, the new approach overcomes this problem through a pruning procedure based on generalized cross-validation.

Furthermore, the efficacy of the new approach was investigated through a computational experience. Results have shown that our proposal generally outperforms DEA and C^2NLS with respect to the mean squared error (MSE). In this way, we have seen that the improvement ranges from 3.39% to 86.23% (with a mean value of 48.77%) regarding DEA and from 6.80% to 69.50% (with a mean value of 20.22%) in the case of C^2NLS. As for the bias, we note that our proposal systematically underestimates the true frontier and that its value usually decreases when the sample size increases.

The main limitation of the new approach is related to methodological issues. Our model is additive in nature, what can limit its performance when the (unknown) true production function to be estimated presents interactions between the inputs of the problem. However, although our model is additive, which *a priori* represents a weakness, we have shown using the simulation scenarios taken from Kuosmanen and Johnson (2010) that the new method can outperform DEA and C^2NLS even in non-additive situations. Nevertheless, in practice, these results will depend on the type of interaction among variables and the degree of curvature, things that will be unknown by researchers. Therefore, in the case of dealing with real-world databases, we suggest applying several different methodologies to get a battery of results in such a way that researchers may make a robust decision about the estimation of the production function. In this way, our method could be seen as a complement to conventional approaches in the field of technical efficiency measurement. Unlike other methods, our approach is based on machine learning techniques, which could represent a fresh perspective when estimating production functions of firms and public institutions is the focus.

Additionally, our approach allows the estimation of production functions in many different contexts: banking, education, management, public policy and so on. Moreover, rather than estimating a production function, we could use revenue, expressed in monetary terms, as a response variable and the inputs as covariables and apply our approach to determine the corresponding revenue function. Furthermore, the production frontier estimation is a problem closely related to the edge estimation problem (see, for example, Daouia et al. 2016). In this regard, see Korostelev and Tsybakov (1993) for the literature therein on the edge estimation problem within the area of image reconstruction. Other possible applications of our approach appear, for example, in medicine, where the probability of contracting a certain disease depends monotonically on certain variables, or in environmental pollution where monotonicity applies to the ozone level as a function of the inversion base temperature (Croux et al., 2012). Another recent area to be approached through our methodology is the inverse problem for Hamilton-Jacobi equations and semiconcave envelopes (see, for example, Esteve and Zuazua, 2020).

We end this section by introducing several lines of future research. First, we suggest the possibility of extending this novel approach to the context of multi-output production. Another remarkable line of study is related to determining a ranking of importance of covariables to the context of production frontiers. To do this, the approach described in Friedman (1991) based on the "ANOVA decomposition" could be adapted. Additionally, we resorted to the spline basis functions for the construction of the estimator. However, other suitable basis functions could be used as long as they can adapt to the framework of production frontier estimation. In terms of methodological development, there is one point to be considered related to how to determine technical inefficiency through different efficiency measures. Another line of research is linked to solving the weakness of the computational cost associated with the new technique. To do this, we could take advantage of parallel computing to reduce the execution time, especially sensitive in multi-sample scenarios. Moreover, it could be interesting to know if the estimator converges to the true production function when $n \to \infty$ and to extend the model to the case of considering interaction among the inputs. These two lines seem to open up an interesting avenue for researching for the future. Furthermore, the incorporation and treatment of uncertainty in the model could be considered as another possible remarkable research line. In this regard, see a survey of previous contributions on uncertainty and Data Envelopment Analysis in Wen (2015). An obvious research line to be followed would be applying the new approach to real databases in a variety of empirical contexts. Finally, we could also resort to the adaptation of the smoothing procedures introduced by Friedman (1991) to get a smoothed version of the estimate of the production function. This estimate should satisfy the postulates of monotonicity and concavity, something that is not trivial and deserves future study.

Acknowledgments

We thank two anonymous reviewers for providing constructive comments and help in improving the content and presentation of this paper. Additionally, J. Aparicio thanks the grant PID2019-105952GB-I00 funded by Ministerio de Ciencia e Innovación/ Agencia Estatal de Investigación /10.13039/501100011033. V. España thanks the financial support from the Generalitat Valenciana under Grant ACIF/2021/135. Additionally, J. Aparicio thanks the grant PROMETEO/2021/063 funded by the Valencian Community (Spain). Finally, M. Esteve thanks the financial support from the Spanish Ministry of Science, Innovation and Universities under Grant FPU17/05365.

References

Aigner, D., Lovell, C. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics, 6*(1), 21–37.
Aragon, Y., Daouia, A., & Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory, 21*(2), 358–389.
Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641–647.
Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association, 84*(405), 289–294.
Bakin, S., Hegland, M., & Osborne, M. R. (2000). Parallel MARS algorithm based on B-splines. *Computational Statistics, 15*(4), 463–484.
Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078–1092.
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.
Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444.
Chen, V. C., Ruppert, D., & Shoemaker, C. A. (1999). Applying experimental design and regression splines to high-dimensional continuous-state stochastic dynamic programming. *Operations Research, 47*(1), 38–53.
Chen, Y., & Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78*(4), 729–754.
Croux, C., Gijbels, I., & Prosdocimi, I. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics, 68*(1), 31–44.
Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics, 140*(2), 375–400.
Daouia, A., Noh, H., & Park, B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 3–30.
Dellnitz, A. (2022). Big data efficiency analysis: Improved algorithms for data envelopment analysis involving large datasets. *Computers & Operations Research, 137*, Article 105553.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155–161.
Eilers, P. H., & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(6), 637–653.
Esteve, C., & Zuazua, E. (2020). The Inverse Problem for Hamilton–Jacobi Equations and Semiconcave Envelopes. *SIAM Journal on Mathematical Analysis, 52*(6), 5627–5657.
Esteve, M., Aparicio, J., Rabasa, A., & Rodriguez-Sala, J. J. (2020). Efficiency analysis trees: A new methodology for estimating production frontiers through decision trees. *Expert Systems with Applications, 162*, Article 113783.
Esteve, M., Aparicio, J., Rodriguez-Sala, J. J., & Zhu, J. (2023). Random Forests and the measurement of super-efficiency in the context of Free Disposal Hull. *European Journal of Operational Research, 304*(2), 729–744.
Färe, R., & Lovell, C. K. (1978). Measuring the technical efficiency of production. *Journal of Economic Theory, 19*(1), 150–162.
Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 1–67.
Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning (Vol. 1, No. 10)*. Springer.
Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215–223.
Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
Kartal Koc, E., & Bozdogan, H. (2015). Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function. *Machine Learning, 101*(1), 35–58.
Koc, E. K., & Iyigun, C. (2014). Restructuring forward step of MARS algorithm using a new knot selection procedure based on a mapping approach. *Journal of Global Optimization, 60*(1), 79–102.
Korostelev, A., & Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*. Springer.
Kuosmanen, T., & Johnson, A. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research, 58*(1), 149–160.
Kuosmanen, T., & Johnson, A. (2017). Modeling joint production of multiple outputs in StoNED: Directional distance function approach. *European Journal of Operational Research, 262*(2), 792–801.
Mammen, E., & Yu, K. (2007). Additive isotone regression. *Lecture Notes-Monograph Series*, 179–195.
Martinez, D. L., Shih, D. T., Chen, V. C., & Kim, S. B. (2015). A convex version of multivariate adaptive regression splines. *Computational Statistics & Data Analysis, 81*, 89–106.
Meyer, M. C. (2013). Semi-parametric additive constrained regression. *Journal of Nonparametric Statistics, 25*(3), 715–730.
Molinos-Senante, M., Maziotis, A., Sala-Garrido, R., & Mocholi-Arce, M. (2023). Assessing the influence of environmental variables on the performance of water companies: An efficiency analysis tree approach. *Expert Systems with Applications, 212*, Article 118844.
Murat, N. (2021). Outlier detection in statistical modeling via multivariate adaptive regression splines. *Communications in Statistics-Simulation and Computation*, 1–12.
Olesen, O. B., & Ruggiero, J. (2022). The hinging hyperplanes: An alternative nonparametric representation of a production function. *European Journal of Operational Research, 296*(1), 254–266.
Özmen, A., & Weber, G. W. (2014). RMARS: Robustification of multivariate adaptive regression spline under polyhedral uncertainty. Journal of Computational and Applied *Mathematics, 259*, 914–924.
Özmen, A., Kropat, E., & Weber, G. W. (2017). Robust optimization in spline regression models for multi-model regulatory networks under polyhedral uncertainty. *Optimization, 66*(12), 2135–2155.
Özmen, A., Weber, G. W., Batmaz, İ., & Kropat, E. (2011). RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set. *Communications in Nonlinear Science and Numerical Simulation, 16*(12), 4780–4787.
Parmeter, C. F., & Racine, J. S. (2013). Smooth constrained frontier analysis. *Recent advances and future directions in causality, prediction, and specification analysis* (pp. 463–488). Springer.
Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science, 44*(1), 49–61.
Simar, L., & Wilson, P. W. (2000a). A general methodology for bootstrapping in non–parametric frontier models. *Journal of Applied Statistics, 27*(6), 779–802.
Simar, L., & Wilson, P. W. (2000b). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis, 13*(1), 49–78.
Taylan, P., Weber, G. W., Liu, L., & Yerlikaya-Özkurt, F. (2010). On the foundations of parameter estimation for generalized partial linear models with B-splines and continuous optimization. *Computers & Mathematics with Applications, 60*(1), 134–143.
Theussl, S., & Hornik, K. (2019). *Rglpk: R/GNU linear programming kit interface. R package version 0.6-4* https://CRAN.R-project.org/package=Rglpk.
Tsai, J. C., & Chen, V. C. (2005). Flexible and robust implementations of multivariate adaptive regression splines within a wastewater treatment stochastic dynamic program. *Quality and Reliability Engineering International, 21*(7), 689–699.
Tsionas, M. G. (2022). Efficiency estimation using probabilistic regression trees with an application to Chilean manufacturing industries. *International Journal of Production Economics*, Article 108492.
Valero-Carreras, D., Aparicio, J., & Guerrero, N. M. (2021). Support vector frontiers: A new approach for estimating production functions through support vector machines. *Omega, 104*, Article 102490.
Valero-Carreras, D., Aparicio, J., & Guerrero, N. M. (2022). Multi-output support vector frontiers. *Computers & Operations Research, 143*, Article 105765.
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Vidoli, F. (2011). Evaluating the water sector in Italy through a two stage method using the conditional robust nonparametric frontier and multivariate adaptive regression splines. *European Journal of Operational Research, 212*(3), 583–595.
Wang, Y., & Wang, S. (2013). Estimating α-frontier technical efficiency with shape-restricted kernel quantile regression. *Neurocomputing, 101*, 243–251.
Wang, Y., Wang, S., Dang, C., & Ge, W. (2014). Nonparametric quantile frontier estimation under shape restriction. *European Journal of Operational Research, 232*(3), 671–678.
Weber, G. W., Batmaz, I., Köksal, G., Taylan, P., & Yerlikaya-Özkurt, F. (2011). CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering, 20*(3), 371–400.
Wen, M. (2015). *Uncertain data envelopment analysis*. Springer.
Yazici, C., Yerlikaya-Özkurt, F., & Batmaz, İ (2015). A computational approach to nonparametric regression: Bootstrapping CMARS method. *Machine Learning, 101*(1), 211–230.
Zhang, H. (1994). Maximal correlation and adaptive splines. *Technometrics, 36*(2), 196–201.
Zhang, H., & Singer, B. H. (2010). *Recursive partitioning and applications*. Springer Science & Business Media.
Zhu, J. (2020). DEA under big data: Data enabled analytics and network data envelopment analysis. *Annals of Operations Research*, 1–23.
Zhu, Q., Wu, J., & Song, M. (2018). Efficiency evaluation based on data envelopment analysis in the big data context. *Computers & Operations Research, 98*, 291–300.