Nonparametric Machine Learning for Stochastic Frontier Analysis: A Bayesian Additive Regression Tree Approach

Zheng Wei*, Huiyan Sang, Nene Coulibaly

[a]*Department of Mathematics and Statistics, Texas A&M University - Corpus Christi, 6300 Ocean Dr., Corpus Christi, 78412, TX, USA*
[b]*Department of Statistics, Texas A&M University, 3143 TAMU, College Station, 77843, TX, USA*
[c]*Department of Mathematics and Statistics, Texas A&M University - Corpus Christi, 6300 Ocean Dr., Corpus Christi, 78412, TX, USA*

Abstract

The stochastic frontier model is widely used in economics, finance, and management to estimate the production function and efficiency of a firm or industry. In the current literature of stochastic frontier analysis, parametric forms of production functions, such as Cobb-Douglas and Translog, are often assumed a priori without validation, which may suffer from model misspecification and lead to biased estimates of efficiency. To address this issue, a new set of stochastic frontier models is proposed, namely, 1) monBART-SFM that constructs a monotone-constrained nonparametric production function via an extension of the monotone Bayesian Additive Regression Tree (monBART) framework; 2) softBART-SFM that is built upon softBART, a smooth version of BART, where both models offer greater flexibility in modeling complex and nonlinear production functions that may have high dimensional inputs. The performance of the proposed method is illustrated through simulation studies and a real data application.

Keywords:
Stochastic frontier model, Monotone Bayesian additive regression tree model, Nonparametric model, Variable selection

*Corresponding author.
Email address: zheng.wei@tamucc.edu (Zheng Wei)

1. Introduction

The stochastic frontier model (SFM) is a widely used econometric approach to estimate the technical efficiency of a decision-making unit (DMU), e.g., firms, farms, organizations, or any entities that produce output using input factors. SFM assumes that there exists a production frontier that represents the maximum output that a DMU can produce using its inputs. However, due to various factors such as managerial inefficiencies or technological limitations, the actual output produced by the DMU may fall short of the frontier. SFM estimates the discrepancy between the actual and frontier outputs, known as the technical inefficiency while accounting for the random noise in the data. This information can be used to identify the sources of inefficiencies and inform decisions that can improve the DMU's performance. SFM has been widely applied in various fields, including agriculture, healthcare, finance, and manufacturing, education, transports, food industry, economy of development, macroeconomic, to evaluate the technical efficiency of DMUs (Wiboonpongse et al., 2015; Wei et al., 2021a; Moshiri et al., 2010; Chandio et al., 2019; Çalmaşur, 2016; D'Elia and Ferro, 2021).

The original formulation of SFM, proposed independently by Aigner et al. (1977) and Meeusen and van Den Broeck (1977), assumes a priori an explicit parametric functional form of the boundary of the set of feasible production plans. Commonly used functional forms include the Cobb-Douglas, Translog, and quadratic forms, among others. Despite of their popularity, these restrictive parametric assumptions of the production function in SFM may lead to model misspecifications. Indeed, previous studies (see,e.g., Giannakas et al., 2003; Ferrara and Vidoli, 2017) reported that many widely-used conventional functional forms of the frontier are often too restrictive and inappropriate to characterize the actual complex production frontier in real applications. Such issues may introduce substantial bias and lead to misleading conclusions about the associations between inputs and output, and the entire production analysis.

To address these issues, researchers have developed non-parametric and semi-parametric versions of SFM to relax the functional form for the production function. Fan et al. (1996) introduced a two-step pseudo-likelihood procedure via kernel regression for frontier function. Kumbhakar et al. (2007) developed a local maximum likelihood principle approach. A Python package (van Rossum, 1995) for the stochastic nonparametric envelope of data method (StoNED) was proposed by employing the concave regression based

on piecewise linear functions (Kuosmanen, 2008; Kuosmanen and Johnson, 2010; Kuosmanen and Kortelainen, 2012). StoNED provides appealing properties and economic interpretation of the coefficients, but it is computationally intensive, especially with high-dimensional input or output data as highlighted by Lee et al. (2013); Ferrara and Vidoli (2017). Parmeter and Racine (2013) further considered a nonparametric kernel smoothing estimator of the frontier model with a shape constraint on the estimator's derivative, by solving a standard quadratic programming problem, proposed by Kuosmanen and Johnson (2010). Prokhorov et al. (2021) extended it in the context with endogenous regressors. However, as the dimensionality of the input space increases, the proportion of data points within a fixed-radius neighborhood of a point tends to decrease exponentially. This "curse of dimensionality" phenomenon can pose challenges for kernel methods in high-dimensional input problems. Recently, Ferrara and Vidoli (2017) developed a semiparametric approach based on the generalized additive models with the sum of spline-based smooth functions for each input variable (GAM-SFM), which generalized the work by Fan et al. (1996). Although interactions among inputs can be modeled by considering tensor product basis in GAM-SFM, the dimension of tensor basis grows rapidly as the number of inputs unless one specifies which inputs have interactions (Ferrara and Vidoli, 2017, See page 763). Simar and Wilson (2022) proposed a transformed model with the nonparametric regression estimator (e.g., kernel local polynomial or orthogonal series estimators) based on local MLE approach (Simar et al., 2017). However, their method requires careful bandwidth selections and may encounter difficulties when dealing with inputs of higher dimensions.

Bayesian methods have also been used in the context of SFM due to their capability of incorporating prior knowledge and convenience to produce uncertainty measures for inefficiency estimates. Van den Broeck et al. (1994) and Griffin and Steel (2007) developed the Bayesian framework with normal noise and some one-sided distributions for inefficiency terms, and introduced the use of Markov Chain Monte Carlo (MCMC) methods for model estimation. Galán et al. (2014) proposed a Bayesian SFM to capture the unobserved inefficiency heterogeneity. However, these models assume a fully parametric log-linear model for the production function. Tsionas (2022) proposed a regression tree approach for efficiency estimation using a single decision tree to partition the input covariate space into hyper-rectangular-shaped subregions. The model fits a monotone concave function within each subregion. However, this only imposes local monotone and concave constraints, and

the author did not show if this would guarantee that the regression tree is globally monotone and/or concave. Moreover, it is known that a single decision tree is prone to overfitting because it has to grow deeper to capture complex functional relationships (See, e.g., P. 301 in Hastie et al., 2001). Most recently, Tsionas et al. (2023) introduced a Bayesian nonparametric framework that utilizes neural networks to approximate the deterministic inefficiency and model noise with a smooth mixture of normal distributions. The resulting average frontier is a weighted average of the Cobb-Douglas or Translog production functions, where the weight may also be a function of input variables. The shape constraints are imposed by using a rejection sampler in MCMC by verifying the desired constraints on a selected grid of input points, which can be expensive because the number of grid points grows rapidly in high-dimensional input space.

In light of these limitations in existing SFMs, we propose a new Bayesian machine learning non-parametric SFM, called BART-SFM, where the production function is modeled by Bayesian additive regression trees, and the efficiency term is assigned with a truncated normal prior. To the best of our knowledge, our work is the first Bayesian additive regression tree approach in the stochastic frontier literature.

The Bayesian Additive Regression Tree (BART) is a class of machine learning decision tree boosting type of approach to non-parametrically model the complex functional relationships between a response variable and a set of predictor variables (Chipman et al., 2010). It has gained widespread popularity in recent years due to its superior performance in various regression and classification tasks. In the regression setting with normal errors, BART has been proved to have strong theoretical guarantees (Rocková et al., 2020) for approximating high dimensional nonlinear functions with complex interactions and different levels of smoothness.

In the production frontier models, it is often assumed that the input factors have monotonic relationships with the production output (Nicholson and Snyder, 2012; Ferrara and Vidoli, 2017). This motivates us to extend the monotone BART (monBART) model recently proposed in Chipman et al. (2022) for the normal regression case to the context of SFM with non-Gaussian inefficiency terms. To the best of our knowledge, monBART has not been extended to the non-Gaussian case. We develop an efficient Bayesian backfitting algorithm for the posterior inference of the SFM model parameters, taking into account the non-Gaussian efficiency terms. In addition, the algorithm extends the Markov chain Monte Carlo algorithm in Chipman

et al. (2022) when calculating the proposal ratios of trees and sampling the proposed leaf node values. Specifically, due to the monotone constraint, the proposal ratio of trees in monBART involves a double integration, and Chipman et al. (2022) proposed to calculate this term numerically by summing over a grid of values. In this work, we derive the closed-form formulas for proposal ratios of trees and employ the closed form skew-normal distribution to sample the proposed leaf node. The improved Bayesian fitting algorithm leads to enhanced sampling efficiency and speed.

Nevertheless, monBART models are constructed by adding together piecewise constant functions, rendering them less suitable for modelling very smooth production functions or concave production functions, as pointed out by the editors and anonymous referees. This motivates us to also introduce the soft version of BART for the stochastic frontier model (softBART-SFM), where data are assigned to each subregion probabilistically. Linero and Yang (2018) developed the softBART posterior concentration theory and showed that the posterior distribution of the nonparametric function f concentrates around any unknown true $\alpha-$Hölder smooth function at near minimax rate. This theory shows that softBART can adapt to different levels of smoothness (See Theorem 2 in Linero and Yang, 2018), including smooth monotone and concave functions as special cases. Even if the true production function is smooth monotone concave, the posterior distribution of softBART-SFM can approximate it well without imposing shape constraints on the prior model. In fact, there is a debate in the SFM literature regarding the necessity of assuming concavity in the production function, especially for high dimensional problems where the production function can be highly complex. Both Kuosmanen (2001) and Ferrara and Vidoli (2017) underline that "there is no valid justification for assuming production sets to be generally concave." Many non-concave production functions can be found in stochastic frontier literature, see, e.g., (Färe and Svensson, 1980), (Battese et al., 2004), and (O'Donnell et al., 2008), among others. The unconstrained softBART-SFM offers substantial flexibility in modeling the production function. Nevertheless, we remark that BART based SFM can be useful for variable selection, interaction detection, and the generation of partial dependence plots between subsets of covariates and response for model diagnostic, which may guide the model specifications of parametric production functions.

The proposed BART-SFM methods have several other appealing advantages over existing SFM models: 1) the Bayesian paradigm allows for the incorporation of prior information into the stochastic frontier analysis and

hence can be more robust in cases where the sample size is small or the production data is sparse; 2) BART-SFM naturally provides posterior distributions of quantities of interests such as technical efficiencies, in addition to the point estimates; 3) BART-SFM is appealing for handling large and complex input data, similar to other ensemble decision tree models such as Random Forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016).

The rest of the paper is organized as follows. A brief review of SFM is given in Section 2. The proposed BART-SFM method is described in Section 2.1. An in-depth explanation of the prior models is provided in Section 2.2. The posterior computation method is given in Section 3. The softBART-SFM is derived in Section 4. The simulation studies and a real data analysis are conducted in Section 5 and Section 6 to illustrate the practical utility of the proposed method. Section 7 closes the paper with a conclusion and future work.

2. Review of SFM

The parametric Stochastic Frontier Model (SFM) is a powerful and popular tool used throughout the productivity modeling literature. Its original formulation was introduced independently by Aigner et al. (1977) and Meeusen and van Den Broeck (1977). The SFM provides a method of estimating the best production frontier, $f(\cdot)$, of different DMUs that describes the maximum output, Y (e.g., crop yield), given input vector, $\boldsymbol{x} = (x_1, \ldots, x_k)^\top \in \mathcal{R}^k$ (e.g., labor and fertilizer costs per acre). The estimated frontier is used to evaluate the relative efficiency of different DMUs. In general, SFM can be applied to problems for which there is a theoretical maximum (minimum) and the observed counterpart is below (above) this value. The discrepancy between the two values measures the inefficiency of a DMU. See, e.g., Greene (2008), for more detailed overviews of SFM.

The canonical parametric SFM for a cross-section of n DMUs can be expressed as followings

$$Y_i = f(\boldsymbol{x}_i|\boldsymbol{\beta}) + V_i - U_i, \qquad i = 1, \ldots, n,$$

where measurement errors $V_i \sim N(0, \sigma_v^2)$, U_i represents non-negative inefficiency terms often modeled by $U_i \sim HN(0, \sigma_u^2)$, the half-normal distribution with scale parameter σ_u^2, and V_i and U_i are assumed to be independent. $f(\cdot)$ defines the production frontier given input $\boldsymbol{x}$, and in literature, it is often

assumed that $f(\boldsymbol{x}|\boldsymbol{\beta})$ has some parametric form involving parameters $\boldsymbol{\beta}$. For example, assuming Y_i is the log-output, the Cobb-Douglas production function (Zellner et al., 1966) has a constant returns to scale property and takes the form: $f(\boldsymbol{x}_i|\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, where β_k are the elasticities of output with respect to x_k. Then, the SFM can be rewritten as :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + V_i - U_i.$$

The above canonical SFM is estimated by likelihood-based inference and has a straightforward interpretation. However, it comes with very restrictive (log) linear functional assumptions that can lead to severely biased parameter estimations under model misspecification, as evidenced in Ferrara and Vidoli (2017) and our numerical results to be shown in Section 5. To relax these parametric assumptions, below we introduce a Bayesian nonparametric method to more flexibly model the production function $f(\boldsymbol{x}_i)$.

2.1. MonBART-SFM for efficiency analysis

In this section, we introduce the BART-based Stochastic Frontier Model (BART-SFM) in a canonical cross-sectional setting as follows,

$$\begin{aligned} Y_i &= \beta_0 + \sum_{j=1}^{m} g_j(\boldsymbol{x}_i; T_j, M_j) + V_i - U_i, \qquad (1) \\ V_i &\overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2), \qquad U_i \overset{\text{i.i.d.}}{\sim} HN(0, \sigma_u^2). \end{aligned}$$

The production function $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{m} g_j(\boldsymbol{x}_i; T_j, M_j)$ is modeled by a summation of the intercept β_0 and additive decision tree weak learner functions, $g_j(\boldsymbol{x}_i; T_j, M_j)$. Specifically, each weak learner function $g_j(\boldsymbol{x}_i; T_j, M_j)$ is determined by a binary decision tree, denoted as T_j, and a vector of parameters at the leaf nodes of T_j, denoted as M_j.

A binary decision tree consists of both internal nodes and leaf nodes, which defines the decision rules to recursively partition the space of input vector $\boldsymbol{x}$ into non-overlapping hyperrectangular regions. It begins from a root node that represents the entire input data space. Each node is associated with a subset of data, and the internal node is also associated with a decision rule to further split the data according to one of the input variables, x_k, $k = 1, \ldots, p$. The most commonly adopted decision rule is taking the form

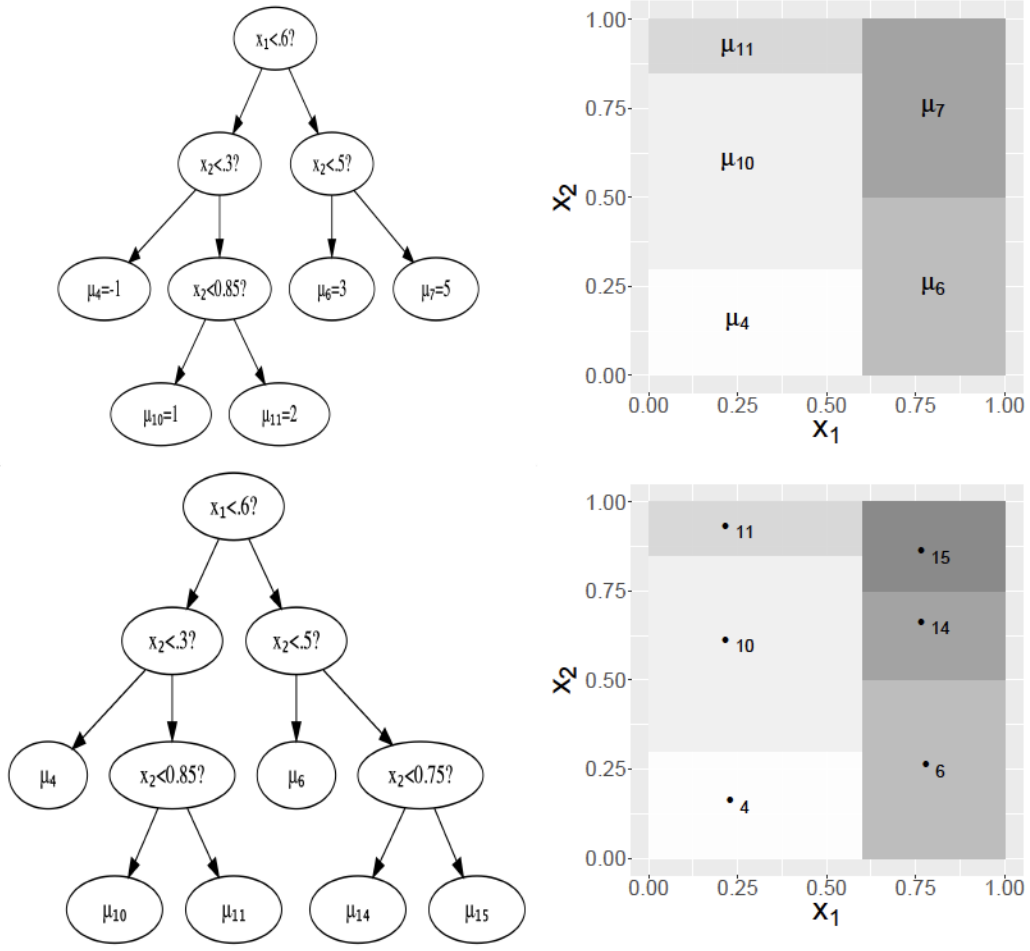

Figure 1: The plot in the top left panel shows a monotone regression tree T with 5 leaf nodes and 4 internal nodes with splitting decision rules. The plot in the bottom left panel shows the proposal tree T^* by further splitting a leaf node μ_7 from T. μ represents the value of each leaf node. The plots on the right panel show the piecewise monotone functions corresponding to T and T^*. The top node is labeled 1 and the ℓth non-leaf node has a left child with label 2ℓ and a right child with label $2\ell + 1$.

$x_k \leq c$ vs $x_k > c$. The decision rule defines a split of an internal node η into two offspring nodes. Accordingly, the subset of data assigned to η is also partitioned into two smaller subsets that are assigned to the two offspring nodes. We call a node terminal or leaf if it does not further split. All leaf nodes form a disjoint partition of the entire input variable space so that for any arbitrary input data point, one can determine the associated leaf node (i.e., the subregion of the partitioned input space) based on the sequence of decision rules from the top to bottom of T_j (see Figure 1).

Given such a partition of the input variable space defined by T_j, data at each leaf node is assumed to take a constant value (also called leaf weight). Assuming there are n_j leaf nodes, we use $M_j = \{\mu_{j1}, \ldots, \mu_{jn_j}\}$ to denote the local leaf weights. With a slight abuse of notation, we also use $\mu_{j\ell}$ to represent the ℓth leaf node of the jth decision tree. As a result, $g_j(\boldsymbol{x}_i; T_j, M_j)$ is modeled by a piece-wise constant function that assigns any input $\boldsymbol{x}$ to a leaf node and assigns the value of $g_j(\boldsymbol{x})$ according to its corresponding leaf weight, i.e.,

$$g_j(\boldsymbol{x}; T_j, M_j) = \sum_{\ell=1}^{n_j} \mu_{\ell j} \mathcal{I}(\boldsymbol{x}; T_j, \ell), \tag{2}$$

where $\mathcal{I}(\boldsymbol{x}; T_j, \ell)$ is the indicator function such that $g(\boldsymbol{x}; T_j, M_j) = \mu_{j\ell}$ if and only if $\boldsymbol{x}$ is associated to ℓth leaf node of tree T_j (see Figure 1).

The task of estimating $g(\boldsymbol{x}_i; T_j, M_j)$ amounts to estimating the decision tree structure T_j and its leaf weights M_j. Following the Bayesian modeling framework, we assume that the decision tree T_j follows a generative tree prior and assign a prior distribution to each leaf weight to model M_j. The detailed prior model specifications will be described in Section 2.2.

In the literature on economic efficiency, it is often reasonable to assume that the production function is a monotone function (Gijbels et al., 1999) because of the free disposability of inputs and outputs property, i.e., an increase in inputs does not decrease production output (Färe et al., 1985; Shephard, 2015). Such an assumption is easily satisfied by (log)linear SFM, but nonparametric SFM models require extra care to guarantee monotonicity. Indeed, in the aforementioned classical BART models, the leaf weights of each decision tree are often assigned with independent priors, which precludes the imposition of a monotonicity constraint on the nonparametric function.

In this paper, we consider a monotone-constrained Bayesian additive decision trees method to model the production function, extending a very re-

cent monotone BART method in Chipman et al. (2022) originally designed for normal-error regression models. We follow the definition of a monotone multivariate function given below,

Definition 2.1. *(Chipman et al., 2022) For a subset $\mathcal{S}$ of the input vector $x \in \mathcal{R}^p$, a function $f : \mathcal{R}^p \to \mathcal{R}$ is said to be monotone in $\mathcal{S}$ if for each $x_k \in \mathcal{S}$, f satisfies*

$$f(x_1, \ldots, x_k + \delta, \ldots, x_p) \geq f(x_1, \ldots, x_k, \ldots, x_p),$$

for all $\delta > 0$ (nondecreasing).

Note that the monotonicity of each decision tree weak learner guarantees the monotonicity of the sum of trees. It suffices to restrict the monotonicity condition for each single tree function $g_j(\boldsymbol{x}; T_j, M_j)$. Recall that each leaf node μ_ℓ of a tree T is represented by a rectangular region of the form

$$R_\ell = \{\boldsymbol{x} | x_k \in [L_{k\ell}, U_{k\ell}), \quad k = 1, \ldots, p\},$$

where the interval $[L_{k\ell}, U_{k\ell})$ for each input x_k is determined by the sequence of splitting rules at internal nodes. R_ℓ is said to be separated from R_{ℓ^*} if $U_{k\ell} < L_{k\ell^*}$, or $L_{k\ell} > U_{k\ell^*}$ for some k. Otherwise, R_ℓ is said to be an above-neighbor of R_{ℓ^*} with respect to x_k if $L_{k\ell} = U_{k\ell^*}$, and below-neighbor of R_{ℓ^*} with respect to x_k if $U_{k\ell} = L_{k\ell^*}$. For example, in T^* of Figure 1, R_{12} is separated from R_4 and R_{11}; R_{12} is the above neighbor of R_7 with respect to X_2, and also the above neighbor of R_{10} with respect to x_1.

The above- and below- neighboring relationships can be used to define a monotone tree by constraining the range of the leaf weight assigned to each leaf node. Specifically, a monotone decision tree is defined by

Definition 2.2. *(Chipman et al., 2022) A tree function $g(\boldsymbol{x}; T, M)$ is monotone in the direction of x_k if the μ level of each of its terminal node regions is:*

(a) not greater than the minimum level among all of its above-neighbor regions with respect to x_k;

(b) not less than the maximum level among all of its below-neighbor regions with respect to x_k.

In addition to the capability of capturing monotonicity, monBART is also shown to provide smoother function estimates, better out-of-sample predictive performance, and tighter posterior credible intervals, compared with standard BART (Chipman et al., 2022). It also inherits several other appealing properties of BART prior models, making it an attractive option for modeling production functions in stochastic frontier models. Firstly, the decision tree-based piece-wise constant function is nonlinear by nature, and an ensemble of many such decision trees can capture complex functional relationships with varying levels of smoothness (Rocková et al., 2020). Secondly, such models are also suitable for handling high-dimensional input variables. We will show in Section 3 that during the construction of the decision tree weak learner, important input variables have a higher chance to be selected to generate internal decision rules, which can help identify the most important predictors in the model and aid in model interpretation. Furthermore, decision trees incorporate decision rules based on various features, enabling the detection of interactions among multiple inputs that may have mixed categorical or numerical types. Additionally, by implementing efficient full-likelihood-based Bayesian inference techniques (to be discussed in Section 3), BART-SFM can handle large datasets and high-dimensional input spaces, while offering a reliable way to estimate model parameters with uncertainty measures.

2.2. Prior Specification

To perform Bayesian inference on BART-SFM, we need to specify the prior distributions for each of the parameters in Eq. (1). The unknown parameters of BART-SFM consist of decision trees $\{T_j\}$, leaf weights $\{M_j\}$, intercept β_0, random error variance σ_v^2 and the variance of the inefficiency term σ_u^2. We assume

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma_u, \sigma_v, \beta_0) = \left[\prod_{j=1}^{m} p(M_j|T_j)p(T_j)\right] p(\sigma_v)p(\sigma_u)p(\beta_0), \quad (3)$$

where the tree components (T_1, M_1), ..., (T_m, M_m) are apriori independent of each other and of σ_v and σ_u, respectively.

For decision trees, we assume that the prior of each T_j follows the generative decision tree prior specified in the following manner:

- Step 1: Start with a root node representing the full input space.

- Step 2: Split a leaf node at tree depth $d = 0, 1, \ldots,$ with probability $p_{\text{split}} = \frac{\alpha}{(1+d)^{\beta}}$. The parameters $\alpha \in (0, 1)$ and $\beta \in [0, \infty)$ are selected to penalize the size of the decision tree by assigning lower splitting probabilities to deeper leaf nodes. This complexity penalization prior encourages shallow decision tree weak learns and aids in preventing BART from overfitting.

- Step 3: If a leaf node splits, with probability p_{rule}, select an input and a split point on that input to split the current leaf node into two new offspring terminal nodes. Uniform distributions are common choices for p_{rule}.

- Apply Steps 2-3 to the two new offspring nodes respectively.

The distribution of the above generative prior of T_j can be written as

$$\pi(T_j) = \prod_{\mu} \left(1 - p_{\text{split}}\left(\mu_j\right)\right) \prod_{\eta} p_{\text{split}}\left(\eta_j\right) \prod_{\eta} p_{\text{rule}}\left(\eta\right),$$

where $\{\mu_j\}$ and $\{\eta_j\}$ represent the leaf and internal nodes of T_j, respectively.

According to the definition of a monotone tree in Def. 2.2, the support of leaf weights needs to be constrained by their neighboring leaf weights. Let C be the set of all trees (T, M) satisfying monotonicity constraints,

$$C = \{(T, M) : g(\boldsymbol{x}; T, M) \text{ is monotone for each } x_k \in \mathcal{S}\}. \tag{4}$$

Then, the prior of leaf weights can be specified by incorporating the above constraints as follows (Chipman et al., 2010):

$$p(M_j|T_j) \propto \left[\prod_{\ell=1}^{n_j} p(\mu_{j\ell}|T_j)\right] \mathcal{I}_C(T_j, M_j), \tag{5}$$

where n_j is the number of leaf nodes of tree T_j and $\mathcal{I}_C$ is the indicator function taking value 1 on C and 0 otherwise.

We now specify the prior for the leaf weights of each M_j given T_j, which is essential to guarantee monotonicity. In standard BART, conjugate normal priors are assigned to $p(\mu_{j\ell}|T_j)$ such that the marginal likelihood ratio for sampling trees and the posterior conditional distributions of leaf weights have closed forms. The hyperparameters in the normal priors are chosen based on the empirical Bayes strategy such that the induced prior on $E(Y|x)$ assigns

a large probability to the empirical range of Y. Similar priors are adopted in monBART but with a correction to account for the monotone constraint. To be specific, for $\mu_{j\ell}$ constrained by C in (4), we assume

$$p(\mu_{ij}|T_j) \sim N(\mu_\mu, c^2\sigma_\mu^2), \text{ with } c^2 = \frac{\pi}{\pi - 1}.$$

This prior choice ensures that the prior variances of the constrained leaf weights in monBART approximately match those of the unconstrained leaf weights in standard BART. Chipman et al. (2022) showed that this strategy helps to balance the prior effects across predictors and facilitates the calibrated specification of σ_μ.

For priors of σ_v and σ_u, we use the conditionally conjugate inverse gamma distributions $\sigma_v^2 \sim IG(\alpha_v, \beta_v)$ and $\sigma_u^2 \sim IG(\alpha_u, \beta_u)$. The hyper-parameters α, β in inverse gamma distributions are specified with a data-informed method that assigns a substantial probability to the entire region of plausible σ values. This method would avoid overconcentration and overdispersion and facilitate the mixing of MCMC. To be specific, a value of df is chosen between 3 and 10, and the value of S is selected so that the qth quantile of prior on σ is located at a prior guess, $\hat{\sigma}$. Note that the residual of BART-SFM is the convolution of a normal random variable V and a half-normal random variable U, which follows the skew-normal distribution. Therefore, to get a reasonable $\hat{\sigma}$, we first fit the skew-normal distribution to data using `selm` function in `sn` package (Bowman and Azzalini, 2021), and apply the estimated scale parameter to $\hat{\sigma}$. Our numerical results indicate that this empirical approach works well in practice.

3. Posterior Computation

3.1. Posterior Sampling

We propose to generate MCMC posterior samples for each parameter based on the Bayesian backfitting algorithm. Since the efficiency terms are of particular interest in SFM, our Bayesian algorithm involves a data augmentation on latent variables $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$. We will show below that this data augmentation strategy also facilitates efficient samplings of other model parameters.

To be specific, we apply the Gibbs sampler to draw samples of each parameter from their posterior conditional distributions derived from,

$$p\big((T_1, M_1), \ldots, (T_m, M_m), \beta_0, \sigma_v^2, \sigma_u^2, \boldsymbol{U}\big|\text{Data}\big).$$

First, the Gibbs sampler draws for each tree (T_j, M_j) for $j = 1, \ldots, m$ given $(\boldsymbol{T}_{-j}, \boldsymbol{M}_{-j}, \beta_0, \sigma_v^2, \sigma_u^2, \boldsymbol{U})$ from the conditional posterior distribution

$$p(T_j, M_j | \boldsymbol{T}_{-j}, \boldsymbol{M}_{-j}, \beta_0, \sigma_v^2, \sigma_u^2, \boldsymbol{U})$$

where $\boldsymbol{M}_{-j} = \{M_r, r \neq j\}$ and $\boldsymbol{T}_{-j} = \{T_r, r \neq j\}$. This step is implemented using the Bayesian backfitting algorithm by sampling each tree (T_j, M_j) sequentially from

$$p(T_j, M_j | \boldsymbol{R}_j, \beta_0, \sigma_v^2, \sigma_u^2, \boldsymbol{U}) \propto p(\boldsymbol{R}_j | T_j, M_j, \beta_0, \sigma_v^2, \sigma_u^2, \boldsymbol{U}) \times p(T_j, M_j)$$

where $\boldsymbol{R}_j = \boldsymbol{Y} - \beta_0 - \sum_{r \neq j} g(X; T_r, M_r) - \boldsymbol{U}$. To simulate a Markov chain move from $(T_j^{(t)}, M_j^{(t)}) \rightarrow (T_j^{(t+1)}, M_j^{(t+1)})$, a localized Metropolis-Hasting algorithm is implemented to accommodate the monotonicity constraint condition given in Eq. (5) (Chipman et al., 2022). First, a candidate proposal tree T^* is generated with probability $q(T_j^{(t)} \rightarrow T_j^*)$, and let $q(T_j^* \rightarrow T_j^{(t)})$ denote the probability of the reversed step. There are two basic local moves in proposing a new tree T^*, namely, the birth move and the death move. Trees in the left panel of Figure 1 illustrate an example of the birth/death proposals, where $T \rightarrow T^*$ is a birth move proposal that splits a randomly selected leaf (e.g., leaf node μ_6) of T into two new leaves by assigning a new splitting rule, and $T^* \rightarrow T$ is a death move proposal by merging two randomly selected sibling leaves to prune the tree. It is easy to see that birth/death moves are reversible and complementary to each other.

Under the proposals $(T_j^{(t)}, M_j^{(t)}) \rightarrow (T_j^*, M_j^*)$ with the above two types of local moves, $M_j^{(t)}$ and M_j^* differ only on those μ's having different ancestries under $T_j^{(t)}$ and T_j^*. We further denotes $M_j^{(t)} = (\mu_{j,same}, \mu_{j,old})$ and $M_j^* = (\mu_{j,same}, \mu_{j,new})$, where $\mu_{j,same}$ corresponds to the unchanged leaf nodes and $\mu_{j,old}$ corresponds to the leaf nodes in T_j that are replaced by $\mu_{j,old}$ in T_j^*. Accordingly, we denote $\boldsymbol{R}_{j,old}$, $\boldsymbol{R}_{j,new}$, and $\boldsymbol{R}_{j,same}$ as the residual data instances assigned to $\mu_{j,old}$, $\mu_{j,new}$, and $\mu_{j,same}$, respectively.

Then, the move $T_j^{(t+1)} = T_j^*$ is accepted with probability

$$\alpha = \min \left\{ \frac{q(T_j^* \rightarrow T_j^{(t)})}{q(T_j^{(t)} \rightarrow T_j^*)} \cdot \frac{p(T_j^{(*)})}{p(T_j^{(t)})} \cdot \frac{p(\boldsymbol{R}_{j,new} | T_j^*, \mu_{j,same})}{p(\boldsymbol{R}_{j,old} | T_j^{(t)}, \mu_{j,same})}, 1 \right\}. \tag{6}$$

The proposal and prior ratios of trees of the first two faction terms in (6) are calculated in a similar way as the standard BART. See the Appendix in

Kapelner and Bleich (2013) for a detailed derivation. However, monBART differs from standard BART in the third marginal likelihood ratio term due to the monotonicity constraint. Let's consider the birth move as an example (the death move would be computed similarly), and omit the weak learner index j for simplicity of notation. In this case, a leaf node μ_{old} in $T^{(t)}$ is split into two offering nodes $\mu_{new} = (\mu_{L}, \mu_{R})$ in T^{*}, with the corresponding local residual data, $\boldsymbol{R}_{new} = (\boldsymbol{R}_{L}, \boldsymbol{R}_{R})$. Note that the BART-SFM assumes independence of residuals conditional on the decision trees T_j and the leaf weights M_j. One needs to calculate the local likelihood $p(\boldsymbol{R}_{L}, \boldsymbol{R}_{R}|T^{*}, \mu_{same})$ in the numerator of the third marginal likelihood ratio term in Eq. (6),

$$p(\boldsymbol{R}_{new} = (\boldsymbol{R}_{L}, \boldsymbol{R}_{R})|T_j^{*}, \mu_{same}) = \iint p(\mu_L, \mu_R|T^{*}, \mu_{same}) p(\boldsymbol{R}_{L}|\mu_{L}) \; (7)$$
$$p(\boldsymbol{R}_{R}|\mu_{R}) d\mu_L d\mu_R,$$

where the monotone constrained prior of the local leaf weights is given by

$$p(\mu_L, \mu_R|T^{*}, \mu_{same}) = 2\phi(\mu_L; \mu_\mu, c\sigma_\mu)\phi(\mu_R; \mu_\mu, c\sigma_\mu) I_{\{\mu_L \leq \mu_R\}},$$

where $\phi(x; \mu, \sigma)$ is the density function of normal random variable with location μ and scale σ.

Let n_L and n_R be the number of observations on the left and right leaf nodes, respectively. With some algebra, $p(\boldsymbol{R}_{L}, \boldsymbol{R}_{R}|T_j^{*}, \mu_{same})$ in Eq. (7) can be simplified to

$$p(\boldsymbol{R}_{L}, \boldsymbol{R}_{R}|T_j^{*}, \mu_{same}) = K\Phi\left(0; \frac{B_L}{A_L} - \frac{B_R}{A_R}, c^2\sigma_\mu^2\sigma_v^2\left(\frac{1}{A_R} + \frac{1}{A_L}\right)\right), \; (8)$$

where $\Phi(\cdot; \mu, \sigma^2)$ is the univariate normal cumulative distribution function with mean μ and variance σ^2, $A_L = \sigma_v^2 + c^2\sigma_\mu^2 n_L$, $B_L = c^2\sigma_\mu^2(\sum_{i=1}^{n_L} R_i^{L2}) + \sigma_v^2\mu_\mu$, and $C_L = c^2\sigma_\mu^2 \sum_{i=1}^{n_L} R_i^{L2} + \sigma_v^2\mu_\mu^2$, and $A_R = \sigma_v^2 + c^2\sigma_\mu^2 n_R$, $B_R = c^2\sigma_\mu^2(\sum_{i=1}^{n_R} R_i^{R2}) + \sigma_v^2\mu_\mu$, $C_R = c^2\sigma_\mu^2 \sum_{i=1}^{n_R} R_i^{R2} + \sigma_v^2\mu_\mu^2$, and

$$K = 2\frac{\frac{\sigma_v^2}{\sqrt{A_L A_R}} \exp\left(-\frac{C_L + C_R - \frac{B_L^2}{A_L} - \frac{B_R^2}{A_R}}{2\sigma_v^2 c^2 \sigma_\mu^2}\right)}{(2\pi\sigma_v^2)^{\frac{n_L + n_R}{2}}}.$$

We defer the detailed derivation to Appendix.

Furthermore, we also derive the close form distribution to draw posterior samples of (μ_L, μ_R). Assuming $T^{(t+1)} = T^*$ is accepted, we sample (μ_L, μ_R) conditional on T^* and other parameters with constraint $\{\mu_L \leq \mu_R\}$ from the bivariate posterior distribution given by,

$$p(\mu_L, \mu_R | \boldsymbol{R}_{_L}, \boldsymbol{R}_{_R}, T^*, \mu_{_{same}}) \propto \phi(\mu_L; \xi_L, \tau_L^2)\phi(\mu_R; \xi_R, \tau_R^2) I_{\{\mu_L \leq \mu_R\}}, \quad (9)$$

whose marginal distributions are

$$p(\mu_L | \boldsymbol{R}_{_L}, \boldsymbol{R}_{_R}, T^*, \mu_{_{same}}) \propto \phi(\mu_L; \xi_L, \tau_L^2)\Phi(-(\mu_L - \xi_L); \xi_L - \xi_R, \tau_R^2), \quad (10)$$

$$p(\mu_R | \boldsymbol{R}_{_L}, \boldsymbol{R}_{_R}, T^*, \mu_{_{same}}) \propto \phi(\mu_R; \xi_R, \tau_R^2)\Phi(\mu_R - \xi_R; \xi_L - \xi_R, \tau_L^2), \quad (11)$$

and the conditional distribution is

$$p(\mu_R | \mu_L, \boldsymbol{R}_{_L}, \boldsymbol{R}_{_R}, T^*, \mu_{_{same}}) \propto \phi(\mu_R; \xi_R, \tau_R^2) I_{\{\mu_L \leq \mu_R\}},$$

where $\xi_L = \frac{B_L}{A_L}$, $\tau_L^2 = \sigma_v^2 c^2 \sigma_\mu^2 / A_L$, $\xi_R = \frac{B_R}{A_R}$, $\tau_R^2 = \sigma_v^2 c^2 \sigma_\mu^2 / A_R$, and A_L, B_L, A_R, B_R are given in Eq. (8). Note that Eq. (10) and (11) are densities of the closed skew-normal distributions (González-Farías et al., 2004). Specifically, we have $\mu_L \sim \text{CSN}_{1,1}(\xi_L, \tau_L^2, -1, \xi_L - \xi_R, \tau_R^2)$, and $\mu_R \sim \text{CSN}_{1,1}(\xi_R, \tau_R^2, 1, \xi_L - \xi_R, \tau_L^2)$, which can be easily sampled, for example, by `R` function "`rcsn`" in '`csn`' package (Dmitry Pavlyuk, 2022). Therefore, to sample (μ_L, μ_R) from (9), we first sample μ_L from (10) and then sample $\mu_R | \mu_L$ from a truncated normal distribution in (12).

Remark: In Chipman et al. (2022), the double integration in $p(\boldsymbol{R}_{_L}, \boldsymbol{R}_{_R} | T_j^*, \mu_{_{same}})$ in Eq. (7) is computed numerically by summing over a grid of μ_L and μ_R values, and (μ_L, μ_R) are drawn using the Metropolis-Hastings algorithm. By using our derived closed-form formula given in Eq. (8), the Metropolis-Hastings step for sampling decision trees and the closed-form Gibbs sampler of (μ_L, μ_R) achieves greater accuracy, efficiency, and speed in comparison to the numerical integration approach proposed by Chipman et al. (2022).

The posterior samples of β_0, σ_v^2, σ_u^2, and $\boldsymbol{U}$ are obtained from the following conditional posterior distributions which all have closed forms thanks to the use of conjugate priors as described in Section 2.2,

$$p(\beta_0 | T_1, \ldots, T_m, M_1, \ldots, M_m, \sigma_v^2, \sigma_u^2, \boldsymbol{U}),$$
$$p(\sigma_v^2 | T_1, \ldots, T_m, M_1, \ldots, M_m, \beta_0, \sigma_u^2, \boldsymbol{U}),$$
$$p(\sigma_u^2 | T_1, \ldots, T_m, M_1, \ldots, M_m, \beta_0, \sigma_v^2, \boldsymbol{U}),$$
and
$$p(\boldsymbol{U} | T_1, \ldots, T_m, M_1, \ldots, M_m, \beta_0, \sigma_v^2, \sigma_u^2).$$

The details of the above formulas on the proposed MCMC algorithm are given in the online supplementary material.

Remark. Similar to other Bayesian nonparametric SFM frameworks presented in Tsionas (2022) and Tsionas et al. (2023), the estimated production is not generally concave. If global concavity is required, one may impose the desired concave constraint on a selected grid of points, and discard the posterior draws by using rejection sampling within MCMC, as implemented by Tsionas et al. (2023). However, it should be noted that the grid of points will expand exponentially as the dimension of input grows, making it computationally very expensive to check for the high dimensional case.

3.2. Posterior inference

We obtain model parameter estimates through posterior inference, by calculating summary statistics of the posterior samples, such as the mean, median, standard deviation, and credible intervals of each parameter. In particular, the estimation of observation-specific technical inefficiency score (TE) plays a crucial role in many empirical implementations of the stochastic frontier model, defined as

$$\mathrm{TE} = E(e^{-U_i}|f(\boldsymbol{x}_i), \sigma_v^2, \sigma_u^2) = \frac{\Phi(\mu_*/\sigma_* - \sigma_*)}{\Phi(\mu_*/\sigma_*)} \exp\left(\frac{1}{2}\sigma_*^2 - \mu_*\right), \tag{12}$$

and $\mu_{i*} = (Y_i - f(\boldsymbol{x}_i))\frac{\sigma_u^2}{\sigma_u^2+\sigma_u^2}$ and $\sigma_* = \sqrt{\frac{\sigma_u^2\sigma_v^2}{\sigma_u^2+\sigma_v^2}}$. In Bayesian SFM literature, to evaluate TE for Bayesian SFM, Zhang (2000) proposed to use the Rao-Blackwell estimator (Gelfand and Smith, 1990), from B posterior samples for σ_v, σ_u and $f(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, the average of conditional posterior means of TE can be obtained by,

$$\hat{\mathrm{TE}}_i = \frac{1}{B}\sum_{b=1}^{B} E(e^{-U_i}|\hat{f}^{(b)}(\boldsymbol{x}_i), \hat{\sigma}_v^{2,(b)}, \hat{\sigma}_u^{2,(b)}).$$

4. softBART-SFM

Linero and Yang (2018) developed the softBART to overcome the lack of smoothness of BART. softBART converts each hard decision tree in Eq. (1) into a soft decision tree. They developed the posterior concentration theory

of softBART and showed that softBART can consistently estimate an unknown function $f_0 \in \mathcal{C}^\alpha([0,1]^p)$ with the posterior contraction rate adapting to the smoothness level $\alpha > 0$ (See Theorem 2 in Linero and Yang, 2018). Here $\mathcal{C}^\alpha([0,1]^p)$ denotes the Hölder space with smoothness index α. Because $\mathcal{C}^\alpha([0,1]^p)$ is a very wide class of functions which include the smooth concave function as a sub-class. Theorem 2 in (Linero and Yang, 2018) shows that softBART can approximate the true function asymptotically well, even without imposing prior concavity assumptions.

For this purpose, we further develop the softBART based SFM (softBART-SFM). Rather than $\boldsymbol{x}$ following a deterministic path down the tree, $\boldsymbol{x}$ in softBART follows a probabilistic path that goes left at branch b with splitting rule $x_j \leq c_b$ with probability,

$$\psi(x; c_b, \kappa_b) = \psi\left(\frac{x - c_b}{\kappa_b}\right),$$

where $\kappa_b > 0$ is a bandwidth parameter controlling the sharpness of the decision. And the model approaches a hard decision tree as κ_b goes to 0, and approaches a constant model as κ_b goes to ∞. A common choice for $\psi(x)$ is the logistic gating function, $\psi(x) = (1 + e^{-x})^{-1}$. By averaging over all possible paths, the jth weak learner function $g_j(\boldsymbol{x}; T_j, M_j)$ in Eq. (1) is given by

$$g_j(\boldsymbol{x}; T_j, M_j) = \sum_{\ell=1}^{n_j} \mu_{\ell_j} \varphi(\boldsymbol{x}; T_j, \ell), \tag{13}$$

with

$$\varphi(\boldsymbol{x}; T, \ell) = \prod_{b \in A(\ell)} \psi(\boldsymbol{x}; c_b, \kappa_b)^{1-R_b} \{1 - \psi(\boldsymbol{x}; c_b, \kappa_b)\}^{R_b},$$

where $A(\ell)$ is the set of ancestor nodes of leaf ℓ and $R_b = 1$ if the path to ℓ goes right at b. The smoothness of softBART is achieved by replacing the piecewise constant weak learner with a locally weighted sum of piecewise constants as shown in (13).

The prior specification in softBART-SFM are mostly the same as monBART-SFM as described in Section 2.2 except that the prior of μ's are independent normals without monotone constraint, see (Linero and Yang, 2018) for details. For the prior of the bandwidths κ_b, we have $\kappa_b \sim Exp(r)$, and set $r = 0.1$ as a default value. For the splitting proportions $\boldsymbol{s} = (s_1, \ldots, s_p)$, the

Dirichlet prior $\boldsymbol{s} \sim \mathcal{D}(a/p^{\xi}, \ldots, a/p^{\xi})$, with $\xi \geq 1$ is applied, (see Linero and Yang, 2018; Linero, 2018).

The posterior sampling algorithm of trees in softBART-SFM still use the back-fitting algorithm with split and merge sampler. However, the acceptance ratio calculations in Eq. (6) are different now due to the change of marginal residual data likelihood. Specifically, the marginal residual data likelihood term now takes the form

$$P(\boldsymbol{R}_j|T_j) = \frac{|2\pi\Omega|^{1/2}}{(2\pi\sigma^2)^{n/2}\left|2\pi\sigma_\mu^2\mathrm{I}\right|^{1/2}} \exp\left(-\frac{\|\boldsymbol{R}_j\|^2}{2\sigma^2} + \frac{1}{2}\widehat{\mu}^\top\Omega^{-1}\widehat{\mu}\right),$$

where

$$\Omega = \left(\frac{\sigma_\mu^2}{m}\mathrm{I} + \Lambda\right)^{-1}, \quad \Lambda = \sum_{i=1}^{n_j}\frac{\boldsymbol{\psi}_i\boldsymbol{\psi}_i^\top}{\sigma_v^2}, \quad \hat{\boldsymbol{\mu}} = \Omega\sum_{i=1}^{n_j}\frac{R_{i,j}\boldsymbol{\psi}_i}{\sigma_v^2},$$
$$\boldsymbol{\psi}_i = (\psi_{i1}, \ldots, \psi_{iL_t})^\top \qquad \text{and} \qquad \psi_{i\ell} = \psi(\boldsymbol{x}_i; T_t, \ell).$$

Accordingly, the conditional distribution of leaf weight M_j given T_j follows multivariate normal distribution $N(\hat{\boldsymbol{\mu}}, \Omega)$.

5. Simulation Study

In this section, we design a simulation study to evaluate the performance of the estimators of the proposed monBART-SFM and softBART-SFM with five simulated data generating processes (DGP): (DGP-1) linear production function case; (DGP-2) nonlinear with exponential production function; (DGP-3) nonlinear production function with multiple covariates; (DGP-4) piecewise nonlinear case; (DGP-5) Translog production function with quadratic and interactions. For each DGP, we simulate $n = 200$ observations and repeat the experiments 100 times. For each experiment, we use $m = 100$ trees, and the hyperparameters are set following the data-driven method described in Section 2.2.

DGP-1. Linear case:

$$Y_i = \beta_0 + \beta_1 X_{1i} + V_i - U_i, \qquad i = 1, \ldots, n,$$

where $\beta_0 = 0.9$, $\beta_1 = 1$, $V_i \sim N(0, \sigma_v^2)$ and $U_i \sim HN(0, \sigma_u^2)$ with $\sigma_v = 1$ and $\sigma_u = 0.5$, and X is simulated from Unif$(1.5, 5)$.

DGP-2. Non-linear case:

$$Y_i = \beta_0 + \beta_1 \exp(X_{1i}) + V_i - U_i, \qquad i = 1, \ldots, n,$$

where $\beta_0 = 10$, $\beta_1 = 5$, $V_i \sim N(0, \sigma_v^2)$ and $U_i \sim HN(0, \sigma_u^2)$ with $\sigma_v = 1$ and $\sigma_u = 0.5$, and X is simulated from Unif$(0, 3)$.

DGP-3. Non-linear case with multiple input variables:

$$Y_i = \beta_0 + \beta_1 \exp(X_{1i}) + \beta_2 \exp(X_{2i}) + \beta_3 \exp(X_{3i}) + \beta_4 \exp(X_{4i}) + V_i - U_i,$$

where $\beta_0 = 10$, $\beta_1 = 5$, and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, $V \sim N(0, \sigma_v^2)$, with $\sigma_v = 1$, and $U \sim HN(0, \sigma_u^2)$ with $\sigma_u = 0.5$, $X_{1i} \sim$ Unif$(.5, 3.5)$, $X_{2i} \sim$ Unif$(1.5, 4.5)$, $X_{3i} \sim$ Unif$(2.5, 5.5)$, $X_4 \sim$ Unif$(3.5, 6.5)$, $X_5 \sim$ Unif$(4.5, 7.5)$.

DGP-4. Piecewise nonlinear case:

$$Y_i = \begin{cases} \beta_0 + \beta_1 \exp(X_{1i}) + V_i - U_i, \\ \beta_0 + 10 + \beta_1 \exp(X_{1i}) + V_i - U_i, \end{cases}$$

where $\beta_0 = 10$, $\beta_1 = 5$, $V \sim N(0, \sigma_v^2)$, with $\sigma_v = 1$, and $U \sim HN(0, \sigma_u^2)$ with $\sigma_u = 0.5$, and $X_{1i} \sim$ Unif$(.5, 3.5)$.

DGP-5. Multivariate case with quadratic terms and interactions between variables (Translog Production Function):

$$\begin{aligned} Y_i \;=\; & \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_{11} X_{1i}^2 + \beta_{22} X_{2i}^2 + \beta_{33} X_{3i}^2 \\ & + \beta_{12} X_{1i} X_{2i} + \beta_{13} X_{1i} X_{3i} + \beta_{23} X_{2i} X_{3i} + V_i - U_i, \end{aligned}$$

where $\beta_0 = 3.5$, $\beta_1 = 1.5$, and $\beta_2 = 0.3$ $\beta_3 = 0.5$, $\beta_{11} = -1.1$, $\beta_{22} = \beta_{33} = 0$, $\beta_{12} = 3.5$, $\beta_{13} = \beta_{23} = 0$, $V \sim N(0, \sigma_v^2)$, with $\sigma_v = 1$, and $U \sim HN(0, \sigma_u^2)$ with $\sigma_u = 0.5$ and $X_{1i} \sim$ Unif$(.5, 3.5)$, $X_{2i} \sim$ Unif$(.5, 3.5)$, $X_{3i} \sim$ Unif$(.5, 3.5)$.

Comparison methods. The inference obtained from BART-SFM is examined and compared with the results of the Bayesian linear SFM model and the Semi-parametric generalized additive model based SFM (GAM-SFM)(Ferrara and Vidoli, 2017). To obtain reasonable Bayesian linear model estimates under DGP2, DGP3, and DGP4, the input data are transformed under the exponential function. The GAM-SFM fits the output variable Y using a sum of smooth functions of covariate input variables,

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \psi_0 + \sum_{j=1}^{k} \psi_j(x_j),$$

where $\psi_j(\cdot)$s are smooth functions. The monotonicity constraint is imposed by means of P-spline and is implemented by the pseudo maximum-likelihood

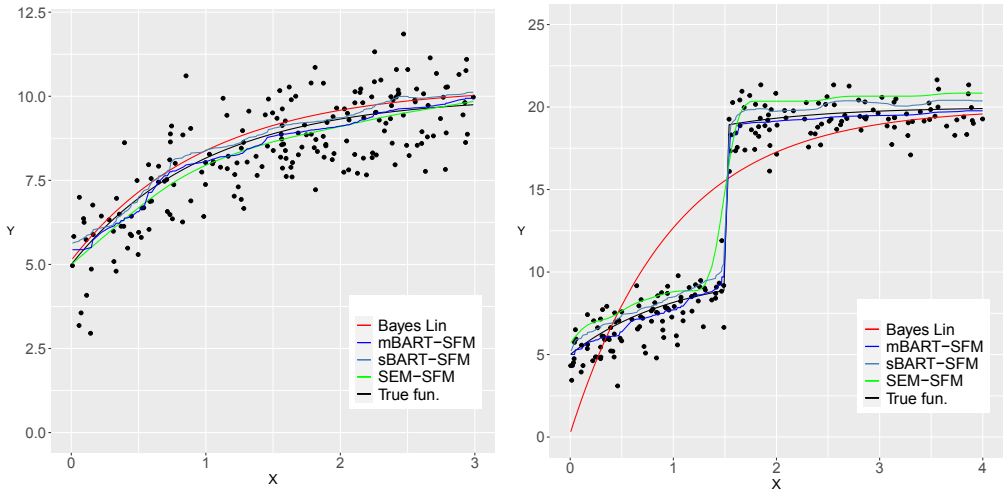

Figure 2: DGP-2 Nonlinear exponential model (left) and DGP-4 Piecewise nonlinear exponential model (right) with the true frontier (black), the GAM-SFA fit(green), Bayesian Linear model fit(red), monBART-SFM fit (blue) and softBART-SFM (light blue).

estimation in the R package `semsfa` (Ferrara and Vidoli, 2021). Note that the package can not model the interactions among covariates (Ferrara and Vidoli, 2017).

Markov chain Monte Carlo implementation. For the Bayesian analysis in the simulation and following real data analyses, we use the *R* programming language for our computations (R Core Team, 2020). B=4,000 MCMC samples are produced for all results. Convergence was assessed using trace plots of σ_u and σ_v and Geweke's convergence diagnostics (Geweke, 1992). The trace plots showed that a burn-in of 1,000 samples was acceptable for all cases and that the chains are mixing well. Note that all of Geweke's convergence diagnostics had values that indicated convergence. The simulation was implemented using R programming language and the computation time for each BART-SFM fitting is about 229 seconds on computers with macOS Big Sur Version 11.6 operating system and Intel Core i7 4.2 GHz Processors with 8 GB memory.

Performance Criteria

To check the performance of the BART-SFM, we report the root mean squared error(RMSE) and absolute value of bias of estimated production function, the average bias of $\hat{\sigma}_v$ and $\hat{\sigma}_u$, and average bias in technical inef-

Table 1: The RMSE, BIAS for SFM based on Bayes Linear model (Lin), GAM-SFM (Sem), BART-SFM (mBT), and softBART-SFM (sBT), $n = 200$ under DGP1 to DGP5 for 100 simulated data sets.

	DGP1				DGP2				DGP3				DGP4				DGP5			
	Lin	Sem	mBT	sBT	Lin	Sem	mBT	sBT	Lin	Sem	mBT	sBT	Lin	Sem	mBT	sBT	Lin	Sem	mBT	sBT
RMSE	16.83	16.09	10.34	14.33	0.05	0.04	0.02	0.03	0.92	0.02	0.02	0.02	0.08	0.03	0.02	0.03	111.33	45.27	24.56	24.45
BIAS	0.57	0.54	0.45	0.53	0.15	0.13	0.11	0.12	0.92	0.11	0.10	0.11	0.22	0.12	0.08	0.11	111.33	0.80	0.57	0.55
Bias_v	0.05	0.09	0.05	0.05	0.07	0.10	0.05	0.05	3.43	0.09	0.04	0.05	1.84	0.17	0.08	0.04	0.31	0.29	0.12	0.06
Bias_u	0.14	0.43	0.39	0.63	0.19	0.44	0.32	0.60	1.62	0.45	0.17	0.64	0.18	0.53	0.40	0.19	0.30	0.54	0.25	0.53
Bias_{TE}	0.06	0.23	0.25	0.07	0.20	0.25	0.07	0.07	0.63	0.21	0.10	0.14	0.23	0.30	0.10	0.30	0.12	0.29	0.15	0.01

ficiency scores. The RMSE and BIAS of estimated production function are defined respectively as,

$$\text{RMSE} = \frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i - Y_i}{Y_i}\right)^2}, \qquad \text{BIAS} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}\left|\frac{\hat{Y}_i - Y_i}{Y_i}\right|. \tag{14}$$

The average bias of $\hat{\sigma}_v$ and $\hat{\sigma}_u$ are given by,

$$\text{Bias}_v = \frac{1}{B}\sum_{b=1}^{B}|\hat{\sigma}_v - \sigma_v|, \qquad \text{Bias}_u = \frac{1}{B}\sum_{b=1}^{B}|\hat{\sigma}_u - \sigma_u|. \tag{15}$$

We also report the mean bias of TE scores obtained by

$$\text{Bias}_{\text{TE}} = \frac{1}{n}\sum_{i=1}^{n}|\hat{\text{TE}}_i - \text{TE}_i|,$$

where TE_i is the true TE score for farm i by using true simulation parameters in Eq. (12).

Figure 2 shows the simulated data under DGP2 with nonlinear patter and DGP4 with piecewise nonlinear pattern. The monBART-SFM (blue curve) and softBART-SFM (light blue curve) fit the data well and are close to the true data-generating function (black curve). The simulation results for DGP1 to DGP5 are summarized in Table 1, from which we have the following observations: (i) The RMSE, BIAS, Bias_v of monBART-SFM and softBART-SFM are all smaller than those in Bayesian linear model and GAM-SFM under DGP1 - DGP5; (ii) The Bias_u of monBART-SFM shows smaller values than GAM-SFM for all cases, and both BART-SFM and GAM-SFM have larger Bias_u than those in the Bayesian Linear model for DGP1-DGP4; (iii) For DGP 5, monBART-SFM shows smallest RMSE, BIAS, Bias_v, Bias_u, and

Bias$_{TE}$ among three models. In summary, monBART-SFM and softBART-SFM shows great flexibility in modeling complex nonlinear patterns and the advantage of fitting data with high dimensional inputs and complex associations. When the true production function is linear, monBART-SFM and softBART-SFM still achieves comparative results as linear models.

6. Real Data Analysis

The World Health Report is a worldwide assessment of the effectiveness of healthcare delivery. In this section, we apply the proposed BART-SFM to the 1996 World Health Report data from 168 countries (WHO, 2000). The output variable Y is a composite measure of health care attainment for each country, and five input variables are considered: health expenditure per capita (X_1), average years of schooling (X_2), Gini coefficient for income inequality (X_3), Gross domestic product (GDP, X_4), and population density (X_5) for each country. The complete panel data set from the year 1993 to 1997 was analyzed using the linear model with the maximum likelihood approach in numerous publications (Greene, 2004, 2005, 2008).

To benchmark the performance of the BART-SFM, we fit linear SFM models using both Bayesian and maximum likelihood approaches. To compare BART-SFM with the Bayesian linear model, the widely applicable information criterion (WAIC) is calculated as the measure of predictive accuracy for models comparison (Gelman et al., 1995). Recall WAIC is defined by

$$\text{WAIC} = -2lppd + 2p_{\text{WAIC}},$$

where $lppd$ denotes the log predictive density and p_{WAIC} is the effective number of parameters to adjust the overfitting,

$$\begin{aligned} lppd &= \sum_{i=1}^{n} \frac{1}{B} \sum_{b=1}^{B} p(y_i|\boldsymbol{\theta}^b), \\ p_{\text{WAIC}} &= 2\sum_{i=1}^{n} \left(\log\left(\frac{1}{B} \sum_{b=1}^{B} p(y_i|\boldsymbol{\theta}^b) \right) - \frac{1}{B} \sum_{b=1}^{B} p(y_i|\boldsymbol{\theta}^b) \right). \end{aligned}$$

The softBART-SFM gives lowest WAIC value (126.25), and the WAIC value (297.84) in monBART-SFM is lower than the WAIC value (342.55) in Bayesian linear model, indicating a better model fit of softBART-SFM potentially due

Table 2: The descriptive statistics of Technical Efficiency for WHO data in 1996 for the SFM with MLE approach (MLE), Bayesian Linear model (Bayes-Lin.) monBART-SFM and softBART-SFM are given in left panel. Five farms with highest efficiency scores and five countries with lowest efficiency scores from softBART-SFM and monBART-SFM are shown in the middle panel and right panel, respectively. Efficiency scores are given in parentheses. The SFM with maximum likelihood approach and Bayesian linear model fails to give the ranks.

					Highest five country code		Least five country code	
	MLE	Bayes-Lin.	sBART	mBART	sBART	mBART	sBART	mBART
Mean	1	0.9953	0.8634	0.9111	167	167	516	828
S.D.	0	0	0.0099	0.0118	(0.8843)	(0.9324)	(0.8164)	(0.8655)
Min	0	0.9952	0.8164	0.8655	608	172	828	516
Max	0	0.9953	0.8843	0.9324	(0.8836)	(0.9303)	(0.8318)	(0.8704)
					172	397	109	109
					(0.8835)	(0.9281)	(0.8391)	(0.8752)
					626	224	430	114
					(0.8826)	(0.928)	(0.8392)	(0.8771)
					562	450	557	547
					(0.8826)	(0.9278)	(0.8445)	(0.8798)

to the more flexible nonparametric production function form. Note that linear model via maximium likelihood approach fails to fit for the WHO data from the year 1996. The "`sfa`" function in '`frontier`' package (Coelli et al., 2013) returns a warning message with model mis-specification and provides zero estimates in MLE of σ_u^2, failing to calculate the technical inefficiency scores. We have similar observations for Bayesian linear model that gives a very small posterior mean for σ_u^2 (0.006). Table 2 shows the summary statistics of the efficiency scores of the fitted models. We observe the range of TE scores in the Bayesian linear model is almost identical to 0.995 due to a small posterior mean in σ_u, while TE scores range from 0.816 upto 0.884 in softBART-SFM and from 0.864 up to 0.931 in monBART-SFM (also see Fig. 3 for empirical density plots of the estimated TE scores).

From the BART-SFM results, we also list countries with the five highest and lowest efficiency scores in Table 2. Furthermore, we can obtain the variable of importance by accounting for the relative frequency of variables used in generating trees from BART-SFM, which are 71.06% for health expenditure per capita and 10.22% for average years of schooling. These indicate the health expenditure per capita and average years of schooling are important factors for the efficiency of the health care delivery system.

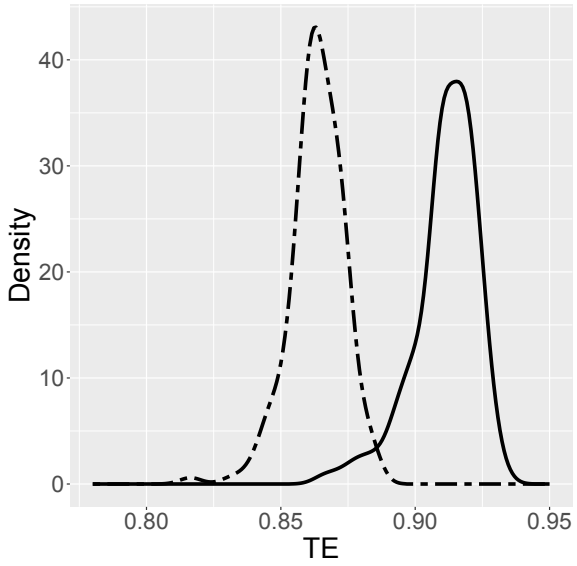

Figure 3: Kernel density of efficiency scores for monBART-SFM (solid line) and softBART (dashed line) using WHO data in 1996. The SFM from the maximum likelihood approach and Bayesian linear model fail to generate the TE scores among countries.

7. Conclusion

In this paper, we propose a Bayesian nonparametric approach for stochastic frontier analysis. The proposed approach is based on the monotone Bayesian Additive Regression Tree framework, which allows greater flexibility in modeling and estimating the production function and technical inefficiencies, while providing uncertainty measures. The method is especially suitable to account for potentially complex relationships with high dimensional data and address variable selections.

The proposed method can be extended in several directions. We considered the case of independent and identically distributed residuals and inefficiency. Nevertheless, in the SFM literature, it is very common to have panel data with time-dependence or data with spatial temporal structure. In future work, the proposed model can be extended to accommodate dependence in such data (Luo et al., 2021, 2022). In addition, extending softBART-SFM models to incorporate shape constraints within their prior models for production functions represents an intriguing yet challenging avenue to explore. Alternatively, neural network based SFM provides another way to impose the concavity on production function by adding concavity constraint on the activation function (Amos et al., 2017). Furthermore, the normal distribution assumption in SFM can be restrictive, which suffers from the "wrong skewness" problem in the literature (Cho and Schmidt, 2020; Wei et al., 2021b). An alternative extension of current work could extend the normal assumption to more flexible distributions in the random noise term in SFM.

Acknowledgement
We thank the editor, the associate editor, and three anonymous referees for their valuable time and suggestions on the paper, which have led to an improved version.

SUPPLEMENTARY MATERIAL

R-package for Bayesian Stochastic Frontier Analysis: R-package BART-SFM containing code to perform the frontier analysis methods described in the article. The package also contains all datasets used as examples in the article. (GNU zipped tar file).

WHO data set: Data set used in the illustration of BART-SFM method in Section 6.

MCMC for BART-SFM fitting: The details on the Bayesian backfitting algorithm for the BART-SFM.

References

Aigner, D., Lovell, C.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. Journal of econometrics 6, 21–37.

Amos, B., Xu, L., Kolter, J.Z., 2017. Input convex neural networks, in: International Conference on Machine Learning, PMLR. pp. 146–155.

Battese, G.E., Rao, D.P., O'donnell, C.J., 2004. A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies. Journal of productivity analysis 21, 91–103.

Bowman, A.W., Azzalini, A., 2021. sn: skew-normal and skew-t distributions. URL: `https://CRAN.R-project.org/package=sn`. r package version 1.0-5.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Van den Broeck, J., Koop, G., Osiewalski, J., Steel, M.F., 1994. Stochastic frontier models: A bayesian perspective. Journal of econometrics 61, 273–303.

Çalmaşur, G., 2016. Technical efficiency analysis in the automotive industry: A stochastic frontier approach. International Journal of Economics, Commerce and Management 4, 120–137.

Chandio, A.A., Jiang, Y., Gessesse, A.T., Dunya, R., 2019. The nexus of agricultural credit, farm size and technical efficiency in sindh, pakistan: A stochastic production frontier approach. Journal of the Saudi Society of Agricultural Sciences 18, 348–354.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. Bart: Bayesian additive regression trees. The Annals of Applied Statistics 4, 266–298.

Chipman, H.A., George, E.I., McCulloch, R.E., Shively, T.S., 2022. mbart: Multidimensional monotone bart. Bayesian Analysis 17, 515–544.

Cho, C.K., Schmidt, P., 2020. The wrong skew problem in stochastic frontier models when inefficiency depends on environmental variables. Empirical Economics 58, 2031–2047.

Coelli, T., Henningsen, A., Henningsen, M.A., 2013. Package 'frontier'.

D'Elia, V., Ferro, G., 2021. Efficiency in public higher education: A stochastic frontier analysis considering heterogeneity. Revista de Análisis Económico–Economic Analysis Review 36, 21–51.

Dmitry Pavlyuk, E.G., 2022. csn: Closed Skew-Normal Distribution. URL: `https://cran.r-project.org/web/packages/csn/`. r package version 1.1.3.

Fan, Y., Li, Q., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. Journal of Business & Economic Statistics 14, 460–468.

Färe, R., Grosskopf, S., Lovell, C.K., 1985. The measurement of efficiency of production. volume 6. Springer Science & Business Media.

Färe, R., Svensson, L., 1980. Congestion of production factors. Econometrica: Journal of the Econometric Society , 1745–1753.

Ferrara, G., Vidoli, F., 2017. Semiparametric stochastic frontier models: A generalized additive model approach. European Journal of Operational Research 258, 761–777.

Ferrara, G., Vidoli, F., 2021. semsfa: Semiparametric estimation of stochastic frontier models. R package version 1.1. Retrieved from `https://CRAN.R-project.org/package=semsfa`.

Galán, J.E., Veiga, H., Wiper, M.P., 2014. Bayesian estimation of inefficiency heterogeneity in stochastic frontier models. Journal of Productivity Analysis 42, 85–101.

Gelfand, A.E., Smith, A.F., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American statistical association 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian data analysis. Chapman and Hall/CRC.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4. Oxford: Oxford University Press, p. 169-193.

Giannakas, K., Tran, K.C., Tzouvelekas, V., 2003. On the choice of functional form in stochastic frontier modeling. Empirical economics 28, 75–100.

Gijbels, I., Mammen, E., Park, B.U., Simar, L., 1999. On estimation of monotone and concave frontier functions. Journal of the American Statistical Association 94, 220–228.

González-Farías, G., Domínguez-Molina, A., Gupta, A.K., 2004. Additive properties of skew normal random vectors. Journal of Statistical Planning and Inference 126, 521–534.

Greene, W., 2004. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the world health organization's panel data on national health care systems. Health economics 13, 959–980.

Greene, W., 2005. Fixed and random effects in stochastic frontier models. Journal of productivity analysis 23, 7–32.

Greene, W.H., 2008. The econometric approach to efficiency analysis. The measurement of productive efficiency and productivity growth 1, 92–250.

Griffin, J.E., Steel, M.F., 2007. Bayesian stochastic frontier analysis using winbugs. Journal of Productivity Analysis 27, 163–176.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2001. The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. Springer-Verlag Inc.

Kapelner, A., Bleich, J., 2013. bartmachine: Machine learning with bayesian additive regression trees. arXiv preprint arXiv:1312.2171 .

Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: a local maximum likelihood approach. Journal of Econometrics 137, 1–27.

Kuosmanen, T., 2001. Dea with efficiency classification preserving conditional convexity. European Journal of Operational Research 132, 326–342.

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

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

Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. Journal of productivity analysis 38, 11–28.

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

Linero, A.R., 2018. Bayesian regression trees for high-dimensional prediction and variable selection. Journal of the American Statistical Association 113, 626–636.

Linero, A.R., Yang, Y., 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. Journal of the Royal Statistical Society Series B: Statistical Methodology 80, 1087–1110.

Luo, Z.T., Sang, H., Mallick, B., 2021. Bast: Bayesian additive regression spanning trees for complex constrained domain. Advances in Neural Information Processing Systems 34, 90–102.

Luo, Z.T., Sang, H., Mallick, B., 2022. Bamdt: Bayesian additive semi-multivariate decision trees for nonparametric regression, in: International Conference on Machine Learning, PMLR. pp. 14509–14526.

Meeusen, W., van Den Broeck, J., 1977. Efficiency estimation from cobb-douglas production functions with composed error. International economic review , 435–444.

Moshiri, H., Aljunid, S.M., Amin, R.M., 2010. Hospital efficiency: Concept, measurement techniques and review of hospital efficiency studies. Malaysian Journal of Public health medicine 10, 35–43.

Nicholson, W., Snyder, C.M., 2012. Microeconomic theory: Basic principles and extensions. Cengage Learning.

O'Donnell, C.J., Rao, D.P., Battese, G.E., 2008. Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. Empirical economics 34, 231–255.

Parmeter, C.F., Racine, J.S., 2013. Smooth constrained frontier analysis. Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr , 463–488.

Prokhorov, A., Tran, K.C., Tsionas, M.G., 2021. Estimation of semi-and nonparametric stochastic frontier models with endogenous regressors. Empirical Economics 60, 3043–3068.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rocková, V., Van der Pas, S., et al., 2020. Posterior concentration for bayesian regression trees and forests. Annals of Statistics 48, 2108–2131.

van Rossum, G., 1995. Python tutorial. Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI). Amsterdam.

Shephard, R.W., 2015. Theory of cost and production functions. Princeton University Press.

Simar, L., Van Keilegom, I., Zelenyuk, V., 2017. Nonparametric least squares methods for stochastic frontier models. Journal of Productivity Analysis 47, 189–204.

Simar, L., Wilson, P.W., 2022. Nonparametric, stochastic frontier models with multiple inputs and outputs. Journal of Business & Economic Statistics , 1–13.

Tsionas, M., Parmeter, C.F., Zelenyuk, V., 2023. Bayesian artificial neural networks for frontier efficiency analysis. Journal of Econometrics 236, 105491.

Tsionas, M.G., 2022. Efficiency estimation using probabilistic regression trees with an application to chilean manufacturing industries. International Journal of Production Economics , 108492.

Wei, Z., Conlon, E.M., Wang, T., 2021a. Asymmetric dependence in the stochastic frontier model using skew normal copula. International Journal of Approximate Reasoning 128, 56–68.

Wei, Z., Zhu, X., Wang, T., 2021b. The extended skew-normal-based stochastic frontier model with a solution to 'wrong skewness' problem. Statistics 55, 1387–1406.

WHO, 2000. The world health report 2000: health systems: improving performance. World Health Organization.

Wiboonpongse, A., Liu, J., Sriboonchitta, S., Denoeux, T., 2015. Modeling dependence between error components of the stochastic frontier model using copula: application to intercrop coffee production in northern thailand. International Journal of Approximate Reasoning 65, 34–44.

Zellner, A., Kmenta, J., Dreze, J., 1966. Specification and estimation of cobb-douglas production function models. Econometrica: Journal of the Econometric Society , 784–795.

Zhang, X., 2000. A monte carlo study on the finite sample properties of the gibbs sampling method for a stochastic frontier model. Journal of Productivity Analysis 14, 71–83.