Computers and Chemical Engineering 143 (2020) 107103

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

ALVEN: Algebraic learning via elastic net for static and dynamic nonlinear model identification

Weike Sun, Richard D. Braatz*

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 3 May 2020
Revised 5 July 2020
Accepted 12 September 2020
Available online 19 September 2020

Keywords:
Process data analytics
Machine learning
Feature selection
Nonlinear regression
Nonlinear model identification
3D Printing

ABSTRACT

An algorithm is proposed that combines nonlinear feature generation and sparse regression to learn interpretable nonlinear models from noisy and limited data. This Algebraic Learning Via Elastic Net for Static and Dynamic Nonlinear Model Identification algorithm employs automated feature generation including families of ubiquitous chemical and biological nonlinear transformations. ALVEN balances model complexity and prediction accuracy through a two-step feature selection procedure, to produce an interpretable model useful for process applications while avoiding overfitting. The generalization to nonlinear dynamical systems, Dynamic ALVEN, is then described. The model accuracy of the algorithms is compared to well-established machine learning methods for a 3D printer and a chemical reactor.

1. Introduction

Data-driven modeling has long been applied by process control engineers to improve product quality (Chiang et al., 2001; von Stosch et al., 2014), from sensors to individual unit operations to entire manufacturing plants. At the unit operation scale, data-driven models are heavily used in controller design, that is, to compute adjustments in manipulated variables to move operational or quality variables towards desirable values. At the plant scale, data-driven models have long been used to predict the final product quality variables from system inputs. Many textbooks and review articles have been written on data-driven modeling, often by the term *system identification* in the control literature (Zhu et al., 1994; Ljung, 2017). Data-driven methods include ARX, ARMAX, state-space identification, and recursive neural networks, and are available in many software packages including in Matlab (Ljung, 1995). Most systems identification methods produce *dense models*, that is, models in which the predictions are functions of all of the inputs, even if the effects of some of the inputs on the outputs are zero.

Despite significant increases in data from modern instrumentation and major advances in machine learning and data science, there has been limited diffusion of these advanced data analytics methods to real manufacturing processes. One major reason is that real manufacturing processes often involve nonlinear dynamic interactions between the manipulated variables and the output variables but do not have the quality and quantity of data needed by many of the data analytics methods. Also, many manufacturing processes have fewer data samples than predictor variables. The data quality can be low because of sensor bias, drift, and noise. The dynamics and/or multivariable interactions are usually not sufficiently excited during the experiments to construct reliable models using many of the modern machine learning techniques such as deep neural networks. Lastly, interpretability of the model is often desired. Interpretability sheds light on the relationship between the system inputs and outputs, making it easier to understand processes and use of the model for other application purposes. An interpretable model that can be constructed using a limited number of samples and easily generalized to nonlinear dynamical systems has been missing. Such models are only a function of a subset of the potential inputs, and the data science literature refers to such models as being *sparse* or *parsimonious*.

To cope with the aforementioned challenges, Algebraic Learning Via Elastic Net (ALVEN), is proposed for the identification of a nonlinear interpretable model for manufacturing data. ALVEN combines nonlinear feature generation for chemical and biological processes with sparse regression, which can be effective for data with a relatively small number of samples. The two-step sparsity-promoting technique used in ALVEN combines the univariate feature selection with the elastic net sparse regression technique (Zou and Hastie, 2005), which is computationally efficient and is able to select, from a large library, the most informative

* Corresponding author at: 77 Massachusetts Avenue, Room E19-551, Cambridge, MA 02139.
E-mail address: braatz@mit.edu (R.D. Braatz).

https://doi.org/10.1016/j.compchemeng.2020.107103
0098-1354/

linear and nonlinear terms. The resulting final model, unlike black-box models, has interpretability, robustness, better generalization capability, and provides physical/chemical insights of the underlying system. Also, ALVEN is easily extended to nonlinear dynamical system identification by dynamic modification on the transformation matrix. The resulting dynamic ALVEN (DALVEN) model has an interpretable input-output structure while also maintaining high modeling accuracy.

The rest of this article is organized as follows. Section 2 briefly reviews nonlinear model identification and sparse regression. Section 3 describes the core concepts of the ALVEN methodology. Section 4 describes the methodology of DALVEN for nonlinear dynamical system identification. Two case studies are presented in Section 5 to demonstrate the effectiveness of ALVEN and DALVEN, respectively. Finally, the conclusion is in Section 6.

2. Background

Nonlinear regression has been an active research field for decades. For systems without first-principles knowledge, black-box methods have been introduced with high approximation capability, including widely applied techniques such as random forest (RF) regression (Breiman, 2001) and support vector regression (SVR) (Vapnik, 2000), and other nonlinear techniques such as *k*-nearest-neighbor (*k*NN) (Altman, 1992) and neural networks (Jain et al., 1996; Sarle, 1994). These black-box methods require access to massive data sets (i.e., prone to overfitting) and lack interpretability. Moreover, it can be difficult to incorporate those black-box methods with other optimization and controller design applications, not only because of their complexity but also due to the discontinuity in some of the functionalized forms (e.g., RF regression, *k*NN).

The identification of interpretable nonlinear regression models is of high practical and research interest. The combination of nonlinear feature generation with feature selection can provide solutions to this problem. Techniques have been developed for generating parsimonious nonlinear regression models. The first category is based on symbolic regression (Forrest, 1993; Koza, 1992; Schmidt and Lipson, 2009). Symbolic regression is an established method for searching among candidate mathematical expressions while minimizing regression errors. The procedure automatically generates the features as well as the form of nonlinear equations. However, symbolic regression is prone to overfitting unless care is explicitly taken to handle model complexity. The second category is based on optimization, e.g., the ALAMO approach (Wilson and Sahinidis, 2017), to search among pre-selected candidate nonlinear transformations that minimize the objective function (e.g., regression error). ALAMO uses a mixed-integer quadratic formulation with Bayesian information criterion (BIC) as the fitness metric to search through a set of nonlinear functions defined by the user, for example, polynomial transformations. The aforementioned techniques are generally computationally expensive and do not scale well to large manufacturing systems of interests.

The third category builds on the linear sparse regression techniques to select among different nonlinear transformations to build a nonlinear interpretable model, e.g., the SINDy modelling for nonlinear dynamical system identification (Brunton et al., 2016; Kaiser et al., 2018). SINDy utilizes sequentially threshold least squares or LASSO to pick nonlinear transformations from user-provided nonlinear transformations. Besides LASSO, there are other sparse regression techniques and an overview is briefly discussed as follows.

The linear sparse regression is formulated as a linear regression plus a penalty for model complexity,

$$\min_{\boldsymbol{w}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_q^q \tag{1}$$

where λ is a positive penalty coefficient that quantifies the relative tradeoff between the complexity of a model and its training error.

For different q, the penalty norm has different forms. If $q=0$, the penalty norm is called the l_0 pseudonorm and the regression problem is also called best subset selection. The optimization that defines l_0-regularization is NP-hard, inherently combinatorial, and computationally expensive to solve. The tightest convex relaxation of the l_0-pseudonorm penalty is the l_1-norm penalty. Sparse regression with the l_1-norm penalty is called the *least absolute shrinkage and selection operator* (*LASSO*) (Tibshirani, 1996), which also promotes sparsity. LASSO tends to select slightly more variables as compared to best subset selection, but is computationally efficient and has been successful in many applications. However, for datasets with a large number of variables $m_x > N$ or highly correlated variables, LASSO has limited performance. The elastic net (EN) has been proposed to resolve the limitations associated with LASSO (Zou and Hastie, 2005). EN does feature selection and continuous shrinkage simultaneously, enabling variable selection with only limited data and can select groups of correlated variables. EN is formulated as

$$\min_{\boldsymbol{w}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\lambda(\alpha\|\boldsymbol{w}\|_1+\tfrac{1-\alpha}{2}\|\boldsymbol{w}\|_2^2) \tag{2}$$

where α is a scalar between 0 and 1 which specifies the tradeoff between the l_1 and l_2 penalties. The l_2 penalty is exactly the same as used in ridge regression, which is a well-established approach for dealing with multicollinearity in the dataset (Zou and Hastie, 2005).

The combination of l_1 and l_2 penalties imposes both sparsity and grouping effects, which provides stable and automated feature selection with good prediction accuracy. The LARS-EN algorithm is proposed (Zou and Hastie, 2005) to solve the EN efficiently, which is based on the LARS algorithm for LASSO and has computational advantages over other optimization techniques for feature selection. Therefore, EN is adopted in ALVEN, and the detailed algorithm for ALVEN is introduced in the next section.

3. ALVEN

ALVEN is a nonlinear regression model learning methodology that builds interpretable, accurate, and robust models from manufacturing data. ALVEN considers nonlinear systems of the form of additive nonlinear transformations of the input variables,

$$y=\sum_i w_i\phi_i(\boldsymbol{x})+\varepsilon \tag{3}$$

where $\phi(\ \cdot\ )$ is the nonlinear mapping function. The nonlinear mapping function does not have any additional parameters to be estimated. Eq. (3) is linear in the parameters w_i, and the optimization (2) is a convex quadratic program.

Sparsity-promoting techniques are used to select the most informative nonlinear transformations from the nonlinear transformation library. ALVEN requires three steps to construct the model, and the detailed algorithm is described as follows.

Step 1. Given the training data matrix $\boldsymbol{X}$, the predictors are transformed by a set of candidate nonlinear functions to $\Theta(\boldsymbol{X})\in\mathbb{R}^{N\times p}$. The candidate nonlinear functions considered in ALVEN are families of typical nonlinear functions in chemical and biological governing equations. For example, power functions are used to construct models for reaction rate laws and to predict concentration from wavenumbers that deviate from Beers law (Milosevic, 2012). The logarithm transformation is often useful when taking the logarithms of both sides of an equation to transform the nonlinear function to an additive form, such as in expressions for heat and mass transfer coefficients and in rate laws for polymer degradation reaction kinetics (Pielichowski and Njuguna, 2005) or

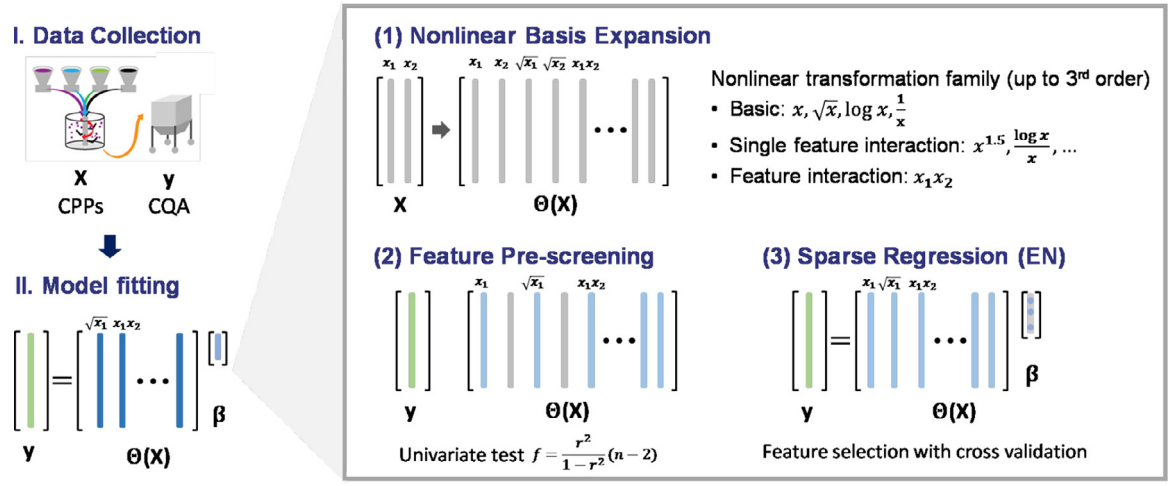

Fig. 1. Schematic of the ALVEN algorithm. The three steps in ALVEN algorithm are (1) nonlinear basis expansion, (2) feature pre-screening via univariate test, and (3) sparse regression via the EN.

the biological growth model. The complexity of the nonlinear transformation is tuned by the degree d, which is up to the 3rd order.

When $d = 1$, the nonlinear transformation family includes basic transformations[1]: x_i, $\sqrt{x_i}$, $\log x_i$, $\frac{1}{x_i}$, $i = 1, \cdots, m_x$;

When $d = 2$, the nonlinear transformation family includes the basic transformation from $d = 1$, and also: (1) 2nd-order interactions between input variables $x_i x_j, \forall i \neq j$ and $i, j = 1, \ldots, m_x$; (2) 2nd-order interactions between basic transformations of each input variable, which include x_i^2, $x_i^{3/2}$, $(\log x_i)^2$, $\frac{\log x_i}{x_i}$, $\frac{1}{x_i^2}$, $x_i^{-1/2}$, $i = 1, \ldots, m_x$;

When $d = 3$, the nonlinear transformation family includes transformations from $d = 2$, and also: (1) 3rd-order interactions between input variables $x_i x_j x_k$, $\forall i \neq j \neq k$ and $i = j \neq k$, $i, j, k = 1, \ldots, m_x$; (2) 3rd-order interactions between basic transformations of each input variable which include x_i^3, $(\log x_i)^3$, $1/x_i^3$, $x_i^{5/2}$, $\frac{(\log x_i)^2}{x_i}$, $\frac{\log x_i}{\sqrt{x_i}}$, $\frac{\log x_i}{x_i^2}$, $x_i^{-3/2}$, $i = 1, \ldots, m_x$.

The final transformed matrix $\Theta(\boldsymbol{X})$ consists of candidate nonlinear transformations of the columns of $\boldsymbol{X}$. For example, when $d = 1$, $\Theta(\boldsymbol{X})$ has the form

$$\Theta(\boldsymbol{X}) = \begin{bmatrix} | & | & | & | \\ \boldsymbol{X} & \sqrt{\boldsymbol{X}} & \log \boldsymbol{X} & 1/\boldsymbol{X} \\ | & | & | & | \end{bmatrix} \tag{4}$$

where the operators $\sqrt{\boldsymbol{X}}$, $\log \boldsymbol{X}$, and $1/\boldsymbol{X}$ refer to the matrix obtained from applying the scalar operation to each element in the matrix

Up to third-order interactions are considered because it is rare for chemical and biological manufacturing systems to have higher order interactions. The transformed nonlinear features may exhibit multicollinearity, which is addressed by the use of elastic net, which employs an l_2-norm to produce models that are interpretable, accurate, and robust (Zou and Hastie, 2005).

Step 2. A pre-screening univariate test is conducted for pre-feature selection in order to accelerate the final feature selection procedure and remedy the curse of dimensionality introduced by the nonlinear basis expansion of the design matrix $\Theta(\boldsymbol{X})$. Here three methodologies are proposed for the pre-screening test.

The default option is to conduct a univariate statistical test for each variable. First, the linear correlation coefficient between the response and each transformed feature $r_i = \text{corr}(y, x_i)$, $\forall x_i \in \Theta(\boldsymbol{x})$ is calculated. Then a univariate statistical test is conducted where the test statistic (scoring function) is calculated from

$$f_i = \frac{r_i^2}{1 - r_i^2}(N - 2) \tag{5}$$

The test statistic follows the $F_{1,N-2}$-distribution and the predictor is retained in the model when the p-value is lower than the predefined significance level, for example, $\alpha = 0.1$.

The second approach is to select the features with the highest scoring function values calculated by Eq. (5) and remove the rest. The percentage of features retained is specified by the user.

The third approach is based on the elbow method (Thorndike, 1953). Basically, after calculating the scoring functions, the values are sorted from the highest to the lowest. Then, the sorted scoring functions are plotted, and the number of features where there is an elbow on the plot is chosen. Although this method is well established, the identification of this break can be ambiguous and only works well when there is a clear elbow in the plot.

Step 3. A sparse model is constructed using EN on the retained features from Step 2. The optimal hyperparameters of EN (λ and α) and the degree of complexity d are selected by cross-validation. EN is effective for feature selection and has low computational cost for large datasets and for a large number of predictors. The overall schematic of ALVEN is illustrated in Fig. 1.

The resulting sparse nonlinear model inherently makes the tradeoff between model complexity and accuracy, avoiding overfitting to the training data, which is especially important for manufacturing data. As in any method, the performance of the ALVEN methodology for sparse nonlinear model construction depends on the data quality, nonlinear function library, and measured system variables. No single method is optimal for all nonlinear regression problems. ALVEN highlights a specific methodology for building an interpretable nonlinear regression model and can serve as a general framework for nonlinear model construction. When no

[1] The term *basic* refers to transformations that are independent of each other, that is, none of the basic transformations of x_i can be implemented by the finite number of applications of the other basic transformations of x_i. The use of basic transformations for $d = 1$ ensures that none of the features are repeated for any higher d.

prior knowledge is available, these default nonlinear transformations are a good starting point for nonlinear model construction for chemical and biological industrial systems, which is missing in other sparse regression-based nonlinear model-building techniques (Brunton et al., 2016; Kaiser et al., 2018). Besides, ALVEN has high flexibility in that the nonlinear transformation library can be altered to the specific needs of a manufacturing system. For example, the feature transformation can be restricted to only polynomial transformations, or specific forms informed by first-principles knowledge. The output variable y can also be transformed, for example, using the inverse or logarithmic transformation, which commonly arise in reaction engineering and transport phenomena. The transformation of the output variable should be done when heteroscedasticity is observed in the residual of the ALVEN model for the original y. After the final ALVEN model is constructed, manufacturers can apply domain knowledge to consider the reasonableness of the selected model.

While ALVEN provides accurate and robust prediction results, whether ALVEN identifies the model that truly captures the physics of the underlying process depends on the specific application. A simple example of ALVEN revealing the process physics is the application to the physical measurements of force F, mass m, and acceleration a. The identified ALVEN model will give $d = 2$ and the true physic law $F = ma$. Similarly, for the measurements of energy E, mass m, and speed of light c, ALVEN will give $d = 3$ and the true physical relationship $E = mc^2$. ALVEN will also find the true physical relationships for transport phenomena problems when the data are written in terms of dimensionless groups and when logarithmic transformations are taken of both sides. For example, relationships involving mass transfer in flow have the form $\mathrm{Nu}_m = c\mathrm{Re}^a\mathrm{Sc}^b$ and heat transfer in flow have the form $\mathrm{Nu}_h = c\mathrm{Re}^a\mathrm{Pr}^b$. Taking logarithms on both sides results in a linear additive form, and the true physics of the process can therefore be easily determined by ALVEN. This example illustrates how dimensionless groups, which have long been used for constructing data-driven models for chemical systems (Rasmuson et al., 2014), can be used to transform the data before feeding into ALVEN. This transformation is recommended for process modeling in which all of the parameters are known in advance, which is typically true in problems dominated by transport phenomena and simple kinetics. When followed by a logarithmic transformation of input and output variables, ALVEN can reveal the true physical relationship. Regardless of whether the true physical relationship is found, values of input variables that give desired outputs by inserting the model into an optimization problem (Biegler, 2010).

In terms of computational complexity, since ALVEN is based on elastic net, the time complexity of ALVEN is $\mathcal{O}(ALK^2N)$, where A is the grid size of the tuning parameter α in the elastic net, L is the grid size of the tuning parameter λ in the elastic net, K is the number of candidate features, and N is the sample size (Efron et al., 2004).

By only requiring a single convex optimization, ALVEN is much more computationally efficient than best subset selection. By using elastic net, ALVEN shares its higher robustness (Zou and Hastie, 2005) than LASSO (Tibshirani, 1996), which performs poorly for highly correlated variables. By defining features as nonlinear relationships typically observed in chemical and biological systems, ALVEN adds a specificity to problems of industrial importance.

4. DALVEN

ALVEN is designed for the construction of a static nonlinear model. For typical industrial processes with fast sampling rates, a static model might not be sufficient to describe the underlying process. DALVEN extends ALVEN to account for nonlinear dynamic behavior, which provides an interpretable input-output nonlinear dynamic model for the system. DALVEN works in the same way as ALVEN but uses a "time lag shift" method (Ku et al., 1995) to augment the original input space $\boldsymbol{X}$ to form a new expanded space $\hat{\boldsymbol{X}}$, with time-shifted vectors of all the variables in $\boldsymbol{X}$ and $\boldsymbol{y}$.

DALVEN used the form of a nonlinear autoregressive model with exogenous inputs (NARX) structure (Billings, 2013),

$$y_t = \sum_t w_i \phi_i(\boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t-l}, y_{t-1}, \ldots, y_{t-l}) + \varepsilon_t \tag{6}$$

where l is the lag number for the past information. A DALVEN model is constructed in three steps, which are detailed below.

Step 1. There are two options in the DALVEN model for nonlinear mapping. For the first option, the design matrix is augmented with past input and measured output variables with previous l observations,

$$\hat{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{x}_t & \cdots & \boldsymbol{x}_{t-l} & y_{t-1} & \cdots & y_{t-l} \\ \boldsymbol{x}_{t+1} & \cdots & \boldsymbol{x}_{t+1-l} & y_t & \cdots & y_{t+1-l} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_N & \cdots & \boldsymbol{x}_{N-l} & y_{N-1} & \cdots & y_{N-l} \end{bmatrix} \tag{7}$$

Then the nonlinear transformation used in ALVEN (see Section 3) is applied to the lagged design matrix $\hat{\boldsymbol{X}}$. The resulting nonlinear transformation $\Theta(\hat{\boldsymbol{X}})$ has not only the nonlinear mapping of past inputs and outputs but also interactions between inputs and outputs at different time points, for example $y_{t-1}\boldsymbol{x}_{t-2}$. The DALVEN model based on this transformation is denoted as the *DALVEN-full* model.

The second option is to first nonlinearly transform the inputs with the nonlinear mapping in ALVEN, denoted as $\phi(\boldsymbol{x}_t)$. Then the design matrix $\Theta(\hat{\boldsymbol{X}})$ is formulated by augmenting each observation vector with both the previous l observations of transformed system inputs and outputs, and stacking as

$$\Theta(\hat{\boldsymbol{X}}) = \begin{bmatrix} \phi(\boldsymbol{x}_t) & \cdots & \phi(\boldsymbol{x}_{t-l}) & y_{t-1} & \cdots & y_{t-l} \\ \phi(\boldsymbol{x}_{t+1}) & \cdots & \phi(\boldsymbol{x}_{t+1-l}) & y_t & \cdots & y_{t+1-l} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \phi(\boldsymbol{x}_N) & \cdots & \phi(\boldsymbol{x}_{N-l}) & y_{N-1} & \cdots & y_{N-l} \end{bmatrix} \tag{8}$$

This second type of nonlinear transformation allows only nonlinearity and interactions between past inputs variables within the same time point, while the past system outputs are not transformed. The second type provides a more restrictive version of nonlinearity while potentially could provide more robust estimates in specific applications. When the number of training samples is limited or prior knowledge of linearity in outputs is known, it is advised to use the second type of nonlinear transformation.

Step 2. The nonlinear mappings with different lag orders are pre-selected on a pairwise basis. Similar as in Section 3, for each transformed feature in $\Theta(\hat{\boldsymbol{X}})$, the linear correlation coefficient between the response and the transformed feature is calculated and used in the univariate statistical test (see Eq. (5)). Then features with p-values lower than the predefined significance level are retained in the model for model construction. Alternatively, the other two univariate feature screening methods described in Section 3 can be used: the predefined number of features retained and the elbow test.

Step 3. EN is used to construct a sparse model on the retained features from Step 2 to minimize the one-step-ahead prediction error. The optimal hyperparameters are selected based on cross-validation for time series or the information criterion, e.g., Akaike information criterion (AIC)

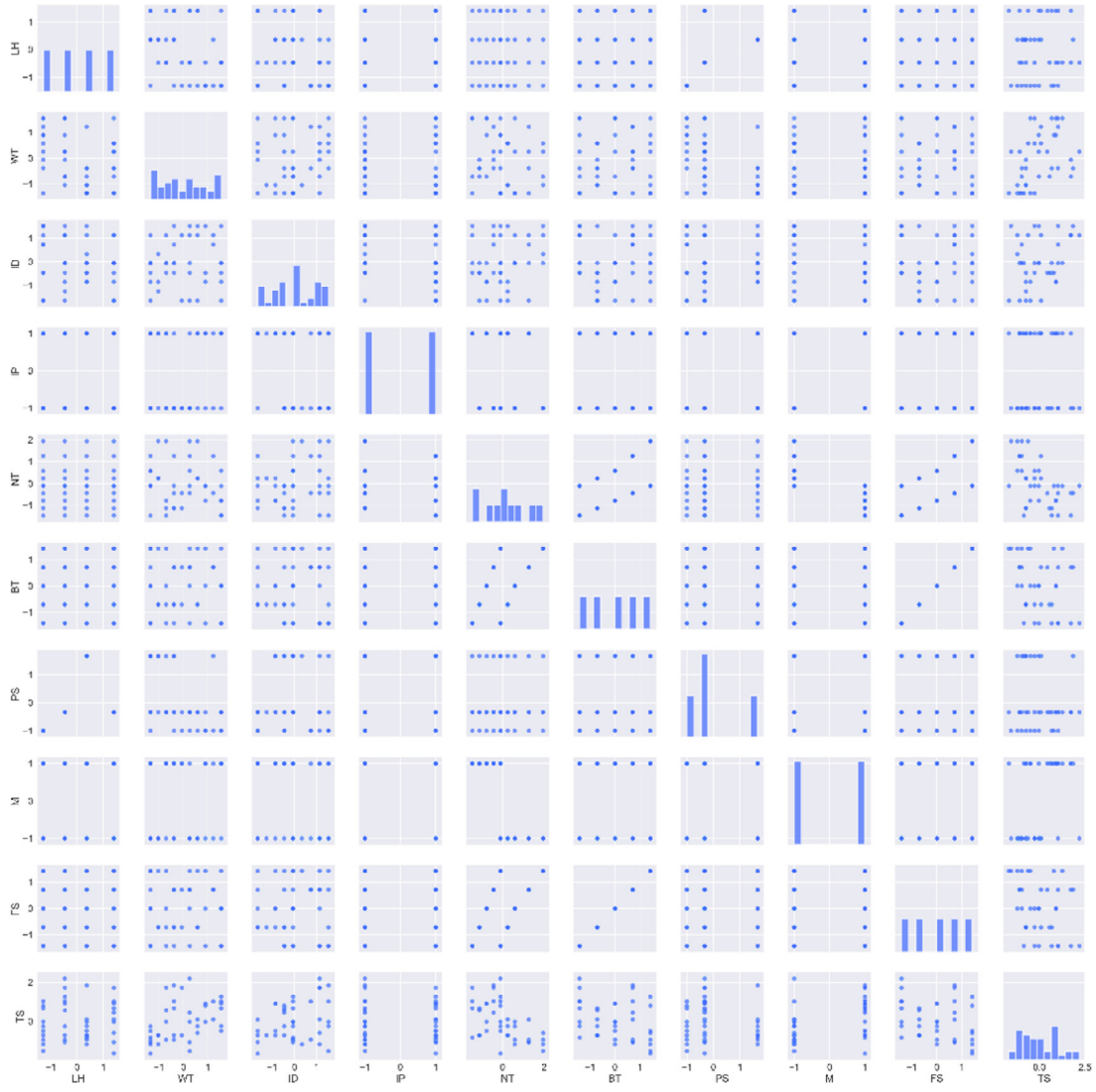

Fig. 2. 3D printer data visualization via single-variable histograms and scatter plots: LH is the layer height, WT is the wall thickness, ID is the infill density, NT is the nozzle temperature, BT is the bed temperature, PS is the print speed, M is the material, FS is the fan speed, and TS is the tension strength.

(Akaike, 1974). After the DALVEN model is fitted, if there is significant autocorrelation remaining, an autoregressive integrated moving average (ARIMA) model can be chosen to construct a time-series model for the residual, which gives DALVEN with ARIMA error. This step can further improve the prediction accuracy of the DALVEN model. The details of the ARIMA procedure are omitted here, and readers can refer to (Chatfield, 1975; Adhikari and Agrawal, 2013) for more information.

The k-step-ahead prediction for DALVEN can be realized through recursion: the output at time t is predicted using previously measured outputs up to time $t-k$ as well as inputs up to time t. For example, a 2-step prediction for the model with lag $l = 1$ is to use $\boldsymbol{x}_t$, $\boldsymbol{x}_{t-1}$, $\boldsymbol{x}_{t-2}$, y_{t-2} to predict y_t, which can be calculated from

$$\hat{y}_t = \sum_i w_i \phi_i(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \hat{y}_{t-1}) \tag{9}$$

where $\hat{y}_{t-1}$ is calculated from

$$\hat{y}_{t-1} = \sum_i w_i \phi_i(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, y_{t-2}). \tag{10}$$

Similar to ALVEN, DALVEN might not have the ability to fit every possible function as compared to some black-box nonlinear dynamic system identification methods, e.g., recurrent neural networks (RNNs). However, DALVEN has a clear input-output nonlinear dynamic structure, which is useful for system understanding and applications such as controller design.

RNNs are widely used to model nonlinear dynamics in the chemical industry, for problems in which the quantity of data is substantial. DALVEN and RNN-based model identification are at opposite ends of the spectrum in terms of constructing discrete-time nonlinear input-output models. RNN-based model identification does not include feature design or feature selection, is formulated in terms of an NP-hard nonconvex optimization, implements a numerical optimization algorithm that does not converge globally, and generates a dense/noninterpretable model. On the other hand, DALVEN extends elasstic net and feature engineering, which are approaches from the machine learning literature. DALVEN includes feature design based on expressions that commonly arise in chemical and biological systems, includes feature selection to generate a sparse/interpretable model, is formulated in terms of a polynomial-time convex optimization, and implements a quadratic optimization that converges globally.

Table 1
Model fitting results for 3D printer data using nested cross-validation.

	Mean			Median			Variance
MSE	Train	Validation	Test	Train	Validation	Test	Test
ALVEN	0.106	0.398	0.322	0.107	0.400	0.298	0.025
RF	0.064	0.602	0.488	0.063	0.601	0.453	0.049
SVR	0.100	0.458	0.456	0.108	0.456	0.422	0.048

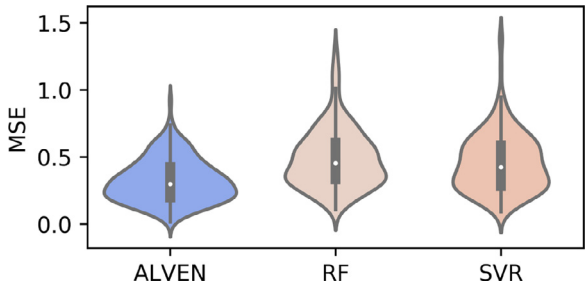

Fig. 3. Testing error distributions via nested cross-validation for 3D printer dataset by three different nonlinear modeling methods.

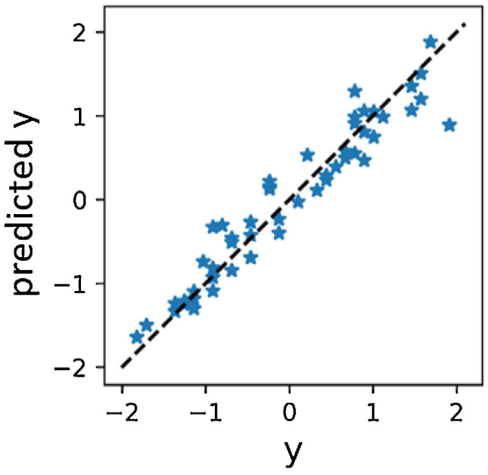

Fig. 4. Final model fitting for 3D printer data using ALVEN.

5. Case studies

In this section, the proposed algorithms are demonstrated to be effective for two process case studies: a dataset collected from a 3D printer for static nonlinear model construction using ALVEN and a simulation dataset of a CSTR for nonlinear dynamical model construction using DALVEN. The proposed methods are compared with other black-box methods to illustrate their effectiveness.

A Python software implementation of the algorithms is available for download (Sun and Braatz, 2020).

5.1. 3D printer

The dataset was collected from samples manufactured by a 3D printer (Ultimaker S5) by a researcher at the TR/Selcuk University (Okudan, 2018). The purpose of the modeling is to relate the adjustment parameters in the 3D printer to the properties of the printed objects.

The dataset has 50 static samples. There are nine input variables (denoted as x_1 to x_9, respectively): layer height (mm), wall thickness (mm), infill density (%), infill pattern, nozzle temperature (°C), bed temperature (°C), print speed (mm/s), material, and fan speed (%). The output variable considered in this case study is the tension strength (MPa) of the printed samples, measured by a professional tension-compression device (a Sincotec GMBH tester), and the test criterion is ASTM d638. The scatter plot of the data is provided in Fig. 2.

ALVEN is compared with RF and SVR for this dataset, as both of these machine learning methods are very well established in the literature. Nested cross-validation is used for model construction and performance evaluation due to the limited number of samples we have. Nested cross-validation provides an unbiased estimation of model performance on unseen data and model stability. In the outer loop of nested cross-validation, 20% of data is used as testing data, and the outer loop is repeated for 180 times. In the inner loop, repeated 3-fold cross-validation with 20 repetitions is implemented. The training, validation, and testing results by three methods are shown in Table 1, and the mean squared error (MSE) distributions for testing datasets over 180 repetitions are shown in Fig. 3. ALVEN is observed to have has much better performance (typically 30% to 50%) over the other methods for this application, in terms of both prediction accuracy (testing data MSE mean/median over 180 repetitions) and model stability (testing data MSE variance over 180 repetitions). Besides accuracy and robustness, ALVEN has interpretability where the nonlinear structure can be obtained from the model.

The final model fitting performance by ALVEN based on all the measured data is shown in Fig. 4, and the residual analysis is shown in Fig. 5. The residual analysis strongly suggests that one of the data points is an outlier, and if the experimentalist was available would be advised to repeat that experiment.

Repeated 3-fold cross-validation with 20 repetitions is used for hyperparameter selection. The final ALVEN model has 55 retained terms, and model coefficients for retained terms are plotted in Fig. 6. The significant terms (with an absolute value above 0.1) are $x_3x_8x_9$, x_1x_3, $x_2x_8x_9$, x_1x_8, and x_5^3. ALVEN provides a direct interpretation of which variables are most informative for predicting the system output and can be easily used for system design (e.g., optimization) purposes. For example, the choice of variables indicate that the tension strength of the 3D-printed product (1) weakly depends on the infill pattern, bed temperature, and print speed, (2) is much more sensitive to nozzle temperature than bed temperature, and (3) depends in a high order on the nozzle temperature.

5.2. CSTR

This dataset is from a simulation of a CSTR where the reaction is exothermic (A $\rightarrow$ B), and the concentration is controlled by regulating the coolant flow (Morningred et al., 1992). The dataset can be downloaded from (De Moor, 2019). The input variable is the coolant flow rate (L/min), while the output variable considered in this example is the concentration of A (mol/L). There are 7500 samples in total: the first 5000 samples are used for training and the rest 2500 samples are used for testing. Fig. 7 is the scatter plot of the data.

Both DALVEN with full nonlinear mapping (denoted as DALVEN-full) and partial nonlinear mapping (denoted as DALVEN), as discussed in Section 4, were applied to the dataset. Both cross-

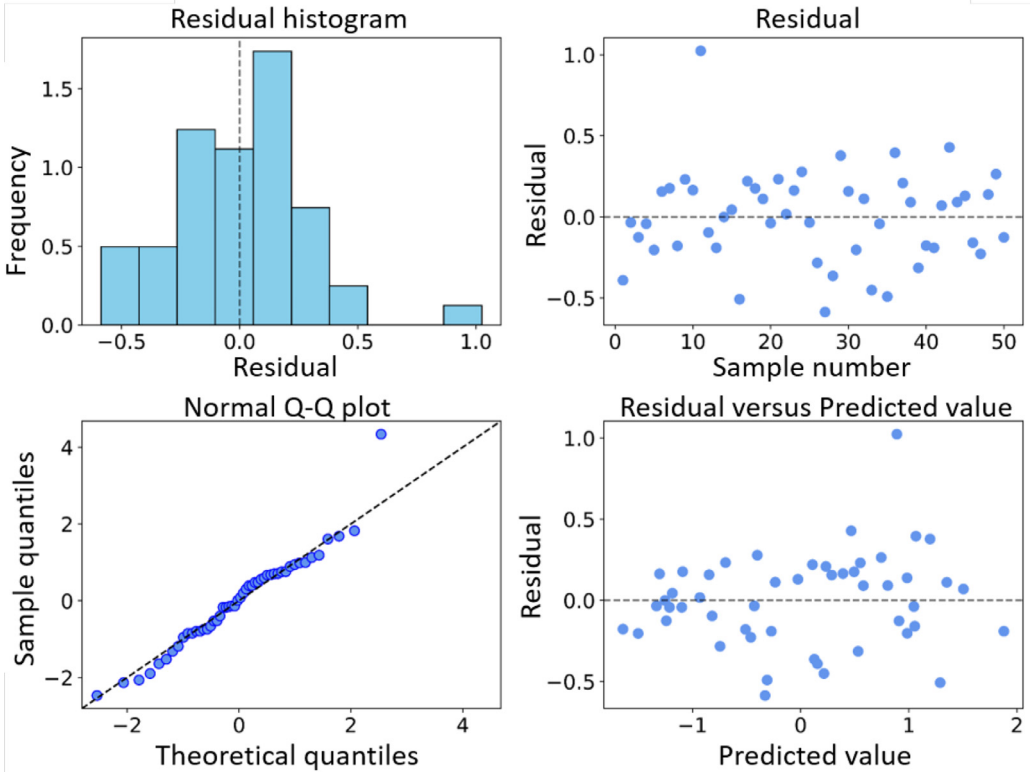

Fig. 5. Residual analysis for ALVEN model.

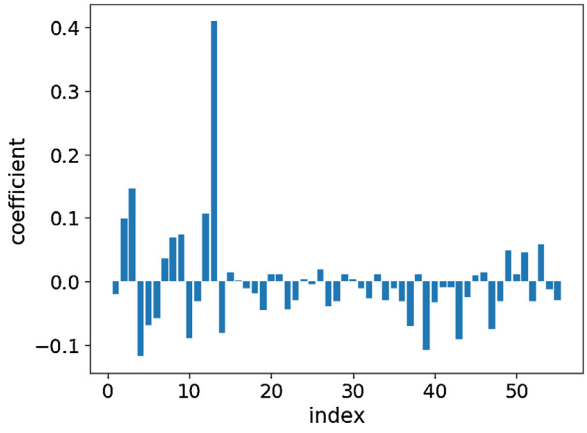

Fig. 6. ALVEN model coefficient magnitudes for the final retained terms.

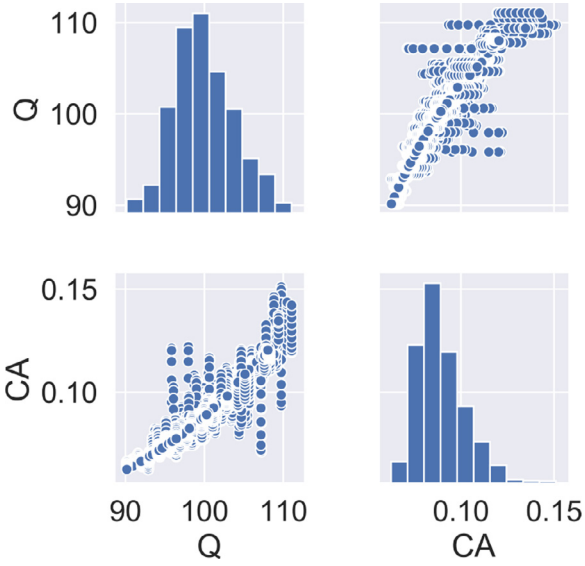

Fig. 7. CSTR data visualization as single-variable histograms and scatter plots: Q is the coolant flow rate and CA is the concentration of A.

validation using a single hold-out dataset (20% of training data) and AIC are used for hyperparameter selections in DALVEN. For comparison, RNNs, with linear, relu, and sigmoid activation functions, are also applied with AIC for hyperparameter selection. RNNs were chosen for being the most popular and widely used machine learning method for the construction of models for nonlinear dynamical systems. During training, the RNN model uses the current input x_t, the previous state $\boldsymbol{s}_t$, and the previously measured output y_{t-1} to predict the next output y_t, which can be stated as

$$\begin{aligned} \boldsymbol{s}_{t+1} &= \phi(\boldsymbol{A}\boldsymbol{s}_t + \boldsymbol{B}\boldsymbol{x}_t + \boldsymbol{D}\hat{\boldsymbol{y}}_t + \boldsymbol{b}) \\ \boldsymbol{y}_t &= \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{k} \end{aligned} \tag{11}$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{k}$ are the parameter matrices and vectors with the appropriate dimension, $\phi(\cdot)$ is the nonlinear activation function, $\boldsymbol{s}$ is the state of the system, and $\hat{\boldsymbol{y}}_{t-1}$ is the actual output measurement at $t-1$. The RNN model is also capable of doing multiple-step-ahead prediction with recursion, and more details can be referred to Sun (2020). The dynamic model performance is compared based on 1-step to 10-step-ahead prediction performance. In the final RNN model selected by AIC, relu is selected as the activation function, and the RNN model has a two-layer architecture with 7 states in each layer.

The training and testing results are shown in Figs. 8 and 9, respectively. DALVEN with full nonlinear mapping gives the best performance for both training and testing, with an order of magnitude better prediction results than RNN for the testing data. RNN

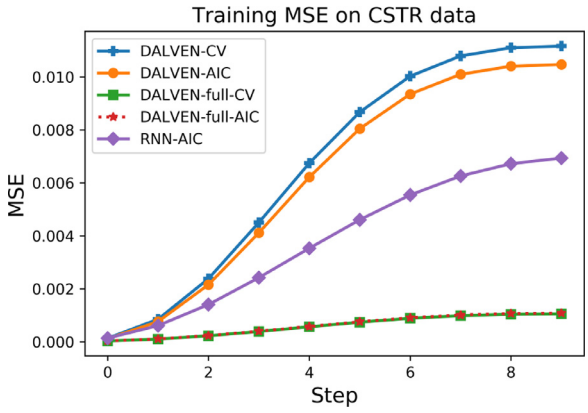

Fig. 8. Training MSE for CSTR data over 10 prediction steps. The values calculated by DALVEN-full-CV and DALVEN-full-AIC are nearly identical, with the largest difference being 0.00003.

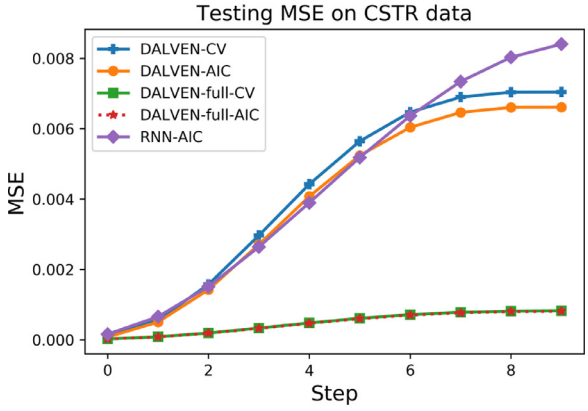

Fig. 9. Testing MSE for CSTR data over 10 prediction steps. The values calculated by DALVEN-full-CV and DALVEN-full-AIC are nearly identical, with the largest difference being 0.00001.

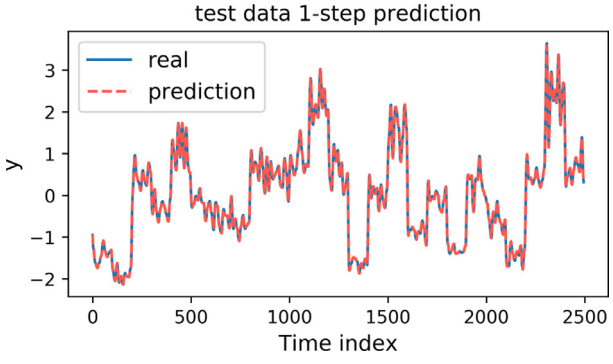

Fig. 10. CSTR 1-step-ahead prediction result for the testing data by DALVEN-full (AIC).

(with relu activation function selected) has lower prediction accuracy and is also hard to interpret. For DALVEN with partial nonlinear mapping, the testing prediction results for long prediction horizon is comparable to RNN, even though its nonlinear form is more restricted than for the RNN.

The testing 1-step prediction by the DALVEN-full model with AIC is plotted in Fig. 10. The final DALVEN-full model degree is 3, the lag order is 4, l_1 ratio is 0.5, and the model has 29 retained terms with interpretable nonlinear forms, and the final model coefficients are plotted in Fig. 11. The significant terms (with absolute values higher than 0.1) are y_{t-1}, $x_t y_{t-1}$, $x_{t-1} y_{t-1}$, and $x_{t-2} y_{t-1}$. These results indicate that the nonlinear dynamics of the CSTR is

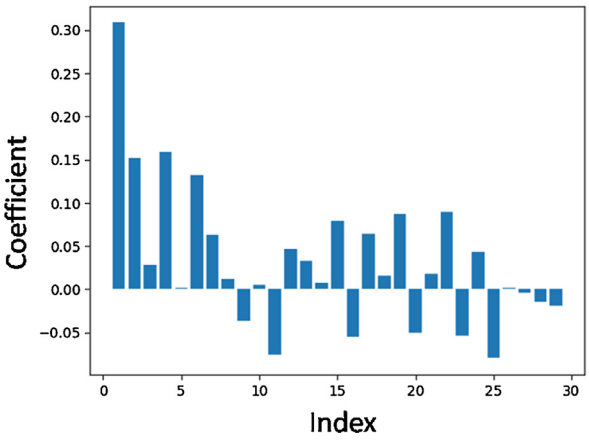

Fig. 11. Model coefficients for the final retained terms in the DALVEN-full (AIC) model of the CSTR.

mostly modeled by only including information of the previous output and the past two values of the predictor. Also, the nonlinearity is mostly bilinear, which can be written as the past output value y_{t-1} multiplied by a linear combination of current and past predictors.

Additional case studies that show qualitatively similar comparative results are available in a Ph.D. thesis (Sun, 2020).

6. Conclusion

This article presents ALVEN, which is an interpretable nonlinear model construction technique to address the problem of learning algebraic functions from manufacturing datasets without information on the nonlinear relationships. ALVEN enables automated feature selection among a family of predefined nonlinear transformations that are suitable for chemical and biological systems. The two-step feature selection procedure, including univariate feature pre-screening and the EN for simultaneous feature post-selection and parameter estimation, enables high model prediction accuracy, robustness, and model interpretability. The proposed algorithm is generalized to dynamic nonlinear system identification, called DALVEN, through an additional step of past information augmentation. ALVEN and DALVEN algorithms are demonstrated on two process case studies and compared with other advanced black-box machine learning algorithms, where both algorithms have shown salient model performance in terms of both model accuracy and interpretability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Weike Sun: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Richard D. Braatz:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Funding acquisition.

Acknowledgement

This work is supported by the U.S. Food and Drug Administration, Grant No. U01FD006483. Any opinions, findings, conclusions,

or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the financial sponsor.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2020.107103.

References

Adhikari, R., Agrawal, R. K., 2013. An introductory study on time series modeling and forecasting. arXiv preprint arXiv:1302.6613.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723. doi:10.1109/TAC.1974.1100705.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. 46 (3), 175–185. doi:10.2307/2685209.

Biegler, L.T., 2010. Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes. SIAM, Philadelphia, Pennsylvania.

Billings, S.A., 2013. Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. John Wiley & Sons, Chichester, U.K..

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. doi:10.1023/A:1010933404324.

Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. 113 (15), 3932–3937. doi:10.1073/pnas.1517384113.

Chatfield, C., 1975. The Analysis of Time Series: Theory and Practice. Springer, Boston, Massachusetts doi:10.1007/978-1-4899-2925-9.

Chiang, L.H., Russell, E.L., Braatz, R.D., 2001. Fault Detection and Diagnosis in Industrial Systems. Springer, London, U.K..

, 2019. DaISy: database for the identification of systems. In: De Moor, B.L.R. (Ed.), Department of Electrical Engineering, ESAT/STADIUS, KU Leuven, Belgium. Used dataset: Continuous stirred tank reactor [98-002], accessed November 30, 2019

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499. doi:10.1214/009053604000000067.

Forrest, S., 1993. Genetic algorithms: principles of natural selection applied to computation. Science 261 (5123), 872–878. doi:10.1126/science.8346439.

Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: a tutorial. Computer 29 (3), 31–44. doi:10.1109/2.485891.

Kaiser, E., Kutz, J.N., Brunton, S.L., 2018. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. Proc. R. Soc. A 474 (2219), 20180335. doi:10.1098/rspa.2018.0335.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, Massachusetts.

Ku, W., Storer, R.H., Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. Chemometr. Intell. Lab. Syst. 30 (1), 179–196. doi:10.1016/0169-7439(95)00076-3.

Ljung, L., 1995. System Identification Toolbox: User's Guide. MathWorks Inc., Natick, Massachusetts.

Ljung, L., 2017. System identification. In: Webster, J.G. (Ed.), Wiley Encyclopedia of Electrical and Electronics Engineering. Wiley Online Library, New York, p. 19.

Milosevic, M., 2012. Internal Reflection and ATR Spectroscopy, vol. 176. John Wiley & Sons, Chichester, U.K..

Morningred, J.D., Paden, B.E., Seborg, D.E., Mellichamp, D.A., 1992. An adaptive nonlinear predictive controller. Chem. Eng. Sci. 47 (4), 755–762. doi:10.1016/0009-2509(92)80266-F.

Okudan, A., 2018. 3D Printer Dataset for Mechanical Engineers. Selçuk University, Konya, Turkey. (accessed November 20, 2019)

Pielichowski, K., Njuguna, J., 2005. Thermal Degradation of Polymeric Materials. iSmithers Rapra Publishing, Crewe, U.K..

Rasmuson, A., Andersson, B., Olsson, L., Andersson, R., 2014. Mathematical Modeling in Chemical Engineering. Cambridge University Press, Cambridge, U.K..

Sarle, W.S., 1994. Neural networks and statistical models. In: Proceedings of the Nineteenth Annual SAS Users Group International Conference, pp. 1538–1550.

Schmidt, M., Lipson, H., 2009. Distilling free-form natural laws from experimental data. Science 324(5923), 81–85. doi:10.1126/science.1165893.

Sun, W., 2020. Advanced Process Data Analytics. Massachusetts Institute of Technology, Cambridge, Massachusetts Ph.D. thesis.

Sun, W., Braatz, R.D., 2020. ALVEN: Algebraic Learning Via Elastic Net: https://github.com/vickysun5/ALVENcode. Massachusetts Institute of Technology, Cambridge, Massachusetts. Software

Thorndike, R.L., 1953. Who belongs in the family? Psychometrika 18 (4), 267–276. doi:10.1007/BF02289263.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer, New York doi:10.1007/978-1-4757-3264-1.

von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S., 2014. Hybrid semi-parametric modeling in process systems engineering: past, present and future. Comput. Chem. Eng. 60, 86–101.

Wilson, Z.T., Sahinidis, N.V., 2017. The ALAMO approach to machine learning. Comput. Chem. Eng. 106, 785–795. doi:10.1016/j.compchemeng.2017.02.010.

Zhu, Y., Van Overschee, P., De Moor, B., Ljung, L., 1994. Comparison of three classes of identification methods. IFAC Proc. Vol. 27 (8), 169–174. doi:10.1016/S1474-6670(17)47710-X.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Seri. B 67 (2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x.