A Kernel Design Approach to Improve Kernel Subspace Identification

Karl Ezra Salgado Pilario, Yi Cao, and Mahmood Shafiee

***Abstract*—Subspace identification methods, such as canonical variate analysis (CVA), are non-iterative tools suitable for the state-space modelling of multi-input, multi-output (MIMO) processes, e.g. industrial processes, using input-output data. To learn nonlinear system behavior, kernel subspace techniques are commonly used. However, the issue of kernel design must be given more attention because the type of kernel can influence the kind of nonlinearities that the model can capture. In this paper, a new kernel design is proposed for CVA based identification, which is a mixture of a global and local kernel to enhance generalization ability and includes a mechanism to vary the influence of each process variable into the model response. During validation, model hyper-parameters were tuned using random search. The overall method is called Feature-Relevant Mixed Kernel Canonical Variate Analysis (FR-MKCVA). Using an evaporator case study, the trained FR-MKCVA models show a better fit to observed data than those of single-kernel CVA, linear CVA, and neural net models under both interpolation and extrapolation scenarios. This work provides a basis for future exploration of deep and diverse kernel designs for system identification.**

***Index Terms*—system identification, kernel PCA, kernel CVA, Newell-Lee evaporator, random search**

I. Introduction

SYSTEM identification, in control literature, refers to the construction of dynamic models from observed input-output process data [1], [2]. Many industrial automation tasks, such as optimization, model predictive control, and fault detection and diagnosis, inherently contain a modelling step. In fact, this step is perceived as a bottleneck, since the quality of an identified model gives an upper bound to the performance of each task [3]. Thus, the need to develop better identification tools for industrial automation remains relevant and worthwhile.

Dynamic systems are typically abstracted as state-space models [4], [5]. The model parameters can then be estimated using either prediction error methods (PEM) or subspace identification methods (SIM) [6]. In PEM, estimation is done by minimizing a cost function of the error between the observed and the predicted outputs. However, because the cost function is usually non-convex, PEMs are trained iteratively and optimality is not guaranteed for large nonlinear systems. In contrast, the more recent SIMs use numerically reliable matrix projections and least-squares regression [6], [7]. This implies that SIMs are non-iterative and are suitable for estimating state-space models for multi-input, multi-output (MIMO) processes, e.g. large industrial plants. Among the SIMs, canonical variate analysis (CVA) was shown to be a reliable method [8], [9], as demonstrated by numerous case studies in chemical processes [10], [11], the Tennessee Eastman Plant [12], and a multiphase flow facility [13], to name a few.

In practice, process systems are inherently nonlinear [3]. Although linear models, such as naive CVA, may still be accurate when working locally around certain operating points, industrial processes nowadays operate at wide-ranging operating conditions, making linear models ineffective. Hence, the state-space approach has been extended to the nonlinear case [3], [14]. Nonlinear PEMs include piecewise affine models [15], recurrent neural networks [16], [17], and neuro-fuzzy methods [18]. Nonlinear SIMs were also developed, mainly by using kernel methods [19]. Here, the idea is to use a kernel function to project nonlinear data onto a high-dimensional feature space where linear methods are applicable. The Gaussian radial basis function (RBF) is often the chosen kernel function so as to impose model smoothness [5], [20], [21]. Early works include the kernel CCA (KCCA) by Lai and Fyfe [22] and Kawahara *et al.* [23]. More recently, KCCA with least-squares support vector machines was presented by Verdult *et al.* [5] and Goethals *et al.* [24]. To avoid overfitting, regularization must be incorporated in SIMs. Regularization via low-dimensional approximation using principal components analysis (PCA) was used in the kernel CVA (KCVA) by Samuel and Cao [20] and in the adaptive KCCA by Van Vaerenbergh *et al.* [21]. Among the SIMs, it is clear that the use of regularized KCCA or KCVA variants is becoming more widely accepted.

However, none of the above kernel subspace methods investigated the impact of kernel design to the model generalization ability. To improve generalization ability, both interpolation

Manuscript received February 20, 2019; revised April 25, 2020; accepted May 14, 2020. This work was supported by the Engineering Research and Development for Technology Program, Department of Science and Technology, Republic of the Philippines. *(Corresponding author: K. E. S. Pilario)*

K. E. S. Pilario is with the Department of Energy and Power, Cranfield University, Cranfield MK43 0AL, United Kingdom, on leave from the Department of Chemical Engineering, University of the Philippines, Diliman, Quezon City 1101, Philippines (e-mail: kspilario@up.edu.ph).

Y. Cao is with the College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: caoyi2018@zju.edu.cn).

M. Shafiee is with the School of Engineering and Digital Arts, University of Kent, Canterbury CT2 7NT, United Kingdom (e-mail: m.shafiee@kent.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier

and extrapolation abilities must be enhanced. Interpolation ability refers to the model accuracy in explaining test data *within* the training region, whereas extrapolation ability refers to the model accuracy in explaining unseen behavior *beyond* the training region. The ability of a model to extrapolate well is important especially when the training data covers only a fraction of the entire operating range of the process, due to safety, time, or cost constraints. However, single-kernel functions, such as the RBF, are known to have limited flexibility [25]–[27]. In other words, it is difficult to simultaneously improve the interpolation and extrapolation abilities of a model by adjusting the parameters of a single kernel only.

In this paper, the main contribution is to address the above mentioned gap by proposing a new kernel design to be applied to regularized kernel CVA. The proposed design involves a convex mixture of global and local kernel types, wherein the latter is built with an ability to determine feature relevance. The idea of mixed kernels have been used in other contexts such as fault detection and diagnosis [27], [28], pattern recognition [26], and time series forecasting [29], [30]. Meanwhile, the feature relevance idea can be traced to the development of the automatic relevance determination kernel for Gaussian process regression [31]. However, their applications to MIMO system identification, via CVA in particular, needs more attention. To this end, we propose a feature-relevant mixed kernel canonical variate analysis (FR-MKCVA) method for nonlinear system identification of MIMO industrial processes. Other major contributions of this work are as follows:

1) A new nonlinear state-space model structure is proposed with FR-MKCVA, which has an output feedback to facilitate multi-step ahead prediction;
2) A methodology to train the nonlinear model is proposed by performing kernel PCA followed by linear CVA;
3) The use of random search [32] is justified for tuning hyper-parameters via hold-out validation; and,
4) The superiority of the tuned FR-MKCVA models to tuned single-kernel CVA, linear CVA, and neural network models is demonstrated using both interpolation and extrapolation tests.

The remainder of this paper is organized as follows: The system identification problem is defined in Section II. The assumed nonlinear model structure is also introduced there. The methodology and the basis for the new kernel design are discussed in Sections III and IV, respectively. In Section V, FR-MKCVA models are evaluated using an industrial case study. Finally, concluding remarks are given in Section VI.

II. Problem Formulation

A. Linear subspace identification

The linear discrete-time stochastic multivariate state-space model can be written as given in [6] as follows:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k \tag{1}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{v}_k \tag{2}$$

where $\mathbf{x}_k \in \mathbb{R}^n$, $\mathbf{u}_k \in \mathbb{R}^{m_u}$, and $\mathbf{y}_k \in \mathbb{R}^{m_y}$ are the state, input, and output column vectors at sampling time k, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m_u}$, $\mathbf{C} \in \mathbb{R}^{m_y\times n}$, and $\mathbf{D} \in \mathbb{R}^{m_y\times m_u}$ are the state-space matrices, and $\mathbf{w}_k \in \mathbb{R}^n$ and $\mathbf{v}_k \in \mathbb{R}^{m_y}$ are independent process and measurement noise at time k, respectively. The identification problem is to find the parameters in the state-space matrices that best explains the response of the system, $\mathbf{y}_k$, when excited by $\mathbf{u}_k$. The SIM approach is to first estimate the state sequences $\mathbf{x}_k$ from the observed $\mathbf{u}_k$ and $\mathbf{y}_k$ data using a suitable subspace method such as CVA. After which, the state-space matrices are obtained by linear regression [6].

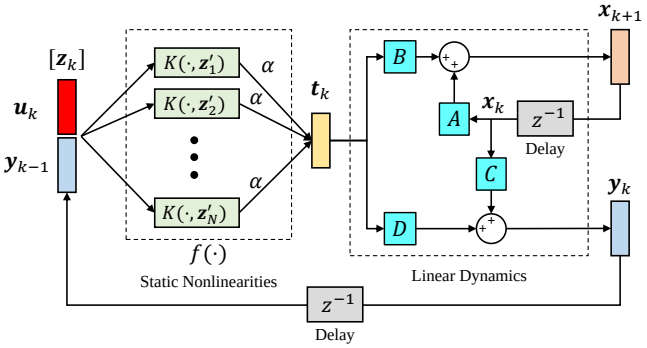


Fig. 1. Block diagram for the proposed nonlinear model structure to be solved in this paper. See Section III for notation details.

B. Nonlinear subspace identification

For the nonlinear case, previous approaches involved kernel CVA [20] or kernel CCA [23] for state reconstruction. Nonlinear regression via neural nets [5] or least-squares support vector machines (LS-SVM) [24], [33] were also used instead of the structure given in (1)-(2).

In this paper, a new stochastic nonlinear state-space structure is proposed as follows:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{t}_k + \mathbf{w}_k \tag{3}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{t}_k + \mathbf{v}_k \tag{4}$$

$$\text{where } \mathbf{t}_k = f(\mathbf{u}_k, \mathbf{y}_{k-1})$$

where inputs and outputs (collectively, $\mathbf{z}_k$) are jointly mapped by a set of nonlinear functions, $f(\cdot)$. The nonlinear functions are specified by the kernel function $K(\cdot\,,\cdot)$, weights α, and the training data set stored in the model as $\mathbf{z}'$. In this structure, the delayed outputs $\mathbf{y}_{k-1}$ are fed back into $f(\cdot)$ so that the model can facilitate multi-step ahead prediction. The new model is depicted in block form in Fig. 1. Note that this structure also arises by replacing the role of $\mathbf{u}_k$ in the original linear model equations (1)-(2) with the new $\mathbf{t}_k$ features.

Following the proposed model, the nonlinear system identification problem is now defined:

'Given N observations of $\mathbf{u}$ and $\mathbf{y}$, determine $f(\cdot)$ and the state-space matrices.'

III. Kernel CVA Methodology

The proposed subspace approach to identify the nonlinear model in (3)-(4) has three steps: (i) unsupervised nonlinear

feature extraction to generate $\mathbf{t}_k$ from the $[\mathbf{u}_k; \mathbf{y}_k]$ data; (ii) state ($\mathbf{x}_k$) estimation using a linear subspace method, and (iii) linear multivariate regression to solve for $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$.

Note that only step (i) is involved with the modelling of system nonlinearities. Nonlinear feature extraction can be achieved by many unsupervised approaches such as neural networks (autoencoders), manifold learning algorithms, etc. This work focuses on kernel methods, the most widely used of which is kernel PCA [34].

A. Kernel PCA

Kernel PCA (KPCA) is an unsupervised learning machine for nonlinear dimensionality reduction [34]. The purpose of KPCA in the overall proposed FR-MKCVA is three-fold: 1) It learns linearly separable features from nonlinear data using kernel projections; 2) It filters noise by discarding the residual subspace of the data; and, 3) It acts as a regularization step to avoid trivial learning in the subsequent linear CVA [26].

Let $\mathbf{z}_k = \begin{bmatrix}\mathbf{u}_k^T & \mathbf{y}_k^T\end{bmatrix}^T \in \mathbb{R}^m$ denote N observations of the input-output training data set, where $k = 1, \ldots, N$ and $m = m_u + m_y$. Initially, all data must be normalized to zero mean and unit variance. In KPCA, the data are first projected from a nonlinear input space onto a high-dimensional feature space F without specifying a mapping $\mathbf{\Phi} : \mathbb{R}^m \to F$ explicitly. Instead, some function can be specified to act as a dot product in F, called the *kernel function* $K(\cdot, \cdot)$. The sample covariance in F can then be computed directly as [34]:

$$\mathbf{K} = [K_{ij}] = [K(\mathbf{z}_i, \mathbf{z}_j)] \in \mathbb{R}^{N \times N}. \tag{5}$$

This covariance matrix, $\mathbf{K}$, is centered to zero-mean by:

$$\widehat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N\mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N\mathbf{K}\mathbf{1}_N \tag{6}$$

where $\mathbf{1}_N \in \mathbb{R}^{N \times N}$ and $(\mathbf{1}_N)_{ij} = 1/N$. To perform PCA, $\widehat{\mathbf{K}}$ is diagonalized as:

$$\widehat{\mathbf{K}}/N = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{7}$$

where $\mathbf{V} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N]$ contains the eigenvectors and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ contains descending eigenvalues. To ensure orthogonality of kernel principal components, the eigenvectors are then scaled to satisfy $\langle \boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i \rangle = 1/\lambda_i$ [35]. To retain relevant information, only the largest ℓ eigenvalues and their respective eigenvectors, i.e. the principal subspace, are collected. The remaining residual subspace may only contain noise, hence, are filtered from the input-output data by choosing $\ell < N$. This step also regularizes the subsequent CVA, wherein ℓ is the regularization parameter [21], [26]. In this work, the number of principal components ℓ is chosen as the first value of i at which $\lambda_i / \operatorname{sum}(\lambda) \leq 5 \times 10^{-5}$.

Taking the first ℓ columns of $\mathbf{V}$, denoted by matrix $\mathbf{V}_\ell$, the projections of the $\mathbf{z}_k$ data are computed as:

$$\mathbf{T} = [\mathbf{t}_k] = \mathbf{V}_\ell^T \widehat{\mathbf{K}} \in \mathbb{R}^{\ell \times N}. \tag{8}$$

Based on (8), it follows that an operator $f(\cdot)$ on any new incoming sample, $\mathbf{z}_k^{\text{new}}$, can be written as:

$$f(\mathbf{z}_k^{\text{new}}) = \left[\sum_{j=1}^{N} \boldsymbol{\alpha}_i^T \widehat{K}(\mathbf{z}_k^{\text{new}}, \mathbf{z}'_j)\right]_{i=1,\ldots,\ell} \tag{9}$$

where $\mathbf{z}'_j$ for $j = 1, \ldots, N$ represents the training data set, and $\widehat{K}$ is the centered kernel function from (6). To achieve good generalization ability, the $f(\cdot)$ must be flexible enough to learn any underlying structure that manifests as system nonlinearities. Upon choosing a kernel function $K(\cdot, \cdot)$ for $f(\cdot)$, a preference for certain classes of structure can be dictated [5], [24], [34]. For instance, a preference for smooth functions is admitted by choosing the well-known Gaussian radial basis function (RBF) kernel.

B. CVA-based States Estimation

In this step, the $\mathbf{t}_k$ features are inputted to linear CVA for estimating the states, $\widehat{\mathbf{x}}_k$, and the model order, n.

Let $\mathbf{t}_k \in \mathbb{R}^\ell$ and $\mathbf{s}_k \in \mathbb{R}^{\ell'}$ represent the features generated from KPCA of $[\mathbf{u}_k\ \mathbf{y}_k]$ and $[\mathbf{y}_k]$ data, respectively. A separate KPCA for $[\mathbf{y}_k]$ is done so that the future inputs $\mathbf{u}_k$ are rendered statistically independent from $\mathbf{t}_k$ when CVA seeks correlations between past $\mathbf{t}_k$ and future $\mathbf{s}_k$ data. Note that the dimensions ℓ and ℓ' can differ in value as they are both obtained using the cumulative percent variance rule from Section III-A.

The past and future data are then collected as follows:

$$\mathbf{Y}_p = \begin{bmatrix} \mathbf{t}_p & \mathbf{t}_{p+1} & \mathbf{t}_{p+2} & \cdots & \mathbf{t}_{p+M-1} \\ \mathbf{t}_{p-1} & \mathbf{t}_p & \mathbf{t}_{p+1} & \cdots & \mathbf{t}_{p+M-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{t}_1 & \mathbf{t}_2 & \mathbf{t}_3 & \cdots & \mathbf{t}_M \end{bmatrix} \tag{10}$$

$$\mathbf{Y}_f = \begin{bmatrix} \mathbf{s}_{p+1} & \mathbf{s}_{p+2} & \mathbf{s}_{p+3} & \cdots & \mathbf{s}_{p+M} \\ \mathbf{s}_{p+2} & \mathbf{s}_{p+3} & \mathbf{s}_{p+4} & \cdots & \mathbf{s}_{p+M+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{s}_{p+f} & \mathbf{s}_{p+f+1} & \mathbf{s}_{p+f+2} & \cdots & \mathbf{s}_N \end{bmatrix} \tag{11}$$

where $\mathbf{Y}_p \in \mathbb{R}^{\ell p \times M}$ and $\mathbf{Y}_f \in \mathbb{R}^{\ell' f \times M}$ are called past and future Hankel matrices, p and f are the lengths of past and future windows, respectively, and $M = N - p - f + 1$ is the maximum number of past-future window pairs that can be taken from the data set. Each ith column of $\mathbf{Y}_p$ and $\mathbf{Y}_f$, denoted respectively as $\mathbf{p}_i$ and $\mathbf{f}_i$, is a pair of data windows at sampling time i, with $i = 1, 2, \ldots, M$. The amount of lags p and f are chosen using autocorrelation analysis [36].

In addition, all windows of length p can be exhausted in an extended past matrix, $\mathbf{Y}_p^+ \in \mathbb{R}^{\ell p \times (N-p+1)}$ [37]:

$$\mathbf{Y}_p^+ = \begin{bmatrix} \mathbf{t}_p & \mathbf{t}_{p+1} & \mathbf{t}_{p+2} & \cdots & \mathbf{t}_N \\ \mathbf{t}_{p-1} & \mathbf{t}_p & \mathbf{t}_{p+1} & \cdots & \mathbf{t}_{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{t}_1 & \mathbf{t}_2 & \mathbf{t}_3 & \cdots & \mathbf{t}_{N-p+1} \end{bmatrix}. \tag{12}$$

Covariance and cross-covariance matrices are computed using the following equations:

$$\mathbf{\Sigma}_{pp} = \mathbf{Y}_p\mathbf{Y}_p^T/(M-1) \in \mathbb{R}^{\ell p \times \ell p} \tag{13}$$

$$\mathbf{\Sigma}_{ff} = \mathbf{Y}_f\mathbf{Y}_f^T/(M-1) \in \mathbb{R}^{\ell' f \times \ell' f} \tag{14}$$

$$\mathbf{\Sigma}_{fp} = \mathbf{Y}_f\mathbf{Y}_p^T/(M-1) \in \mathbb{R}^{\ell' f \times \ell p} \tag{15}$$

CVA aims to find the vectors $\mathbf{a} \in \mathbb{R}^{\ell' f}$ and $\mathbf{b} \in \mathbb{R}^{\ell p}$ so that the correlation between the linear combinations $\mathbf{a}^T\mathbf{f}_i$ and $\mathbf{b}^T\mathbf{p}_i$ is maximized [36]. This correlation is given by:

$$\text{corr}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \boldsymbol{\Sigma}_{fp} \mathbf{b}}{(\mathbf{a}^T \boldsymbol{\Sigma}_{ff} \mathbf{a})^{1/2} (\mathbf{b}^T \boldsymbol{\Sigma}_{pp} \mathbf{b})^{1/2}}. \tag{16}$$

To maximize $\text{corr}(\mathbf{a}, \mathbf{b})$ algebraically, first define $\boldsymbol{\upsilon} = \boldsymbol{\Sigma}_{ff}^{1/2}\mathbf{a}$ and $\boldsymbol{\nu} = \boldsymbol{\Sigma}_{pp}^{1/2}\mathbf{b}$. The CVA problem becomes:

$$\begin{aligned} \max_{\boldsymbol{\upsilon}, \boldsymbol{\nu}} \ & \boldsymbol{\upsilon}^T (\boldsymbol{\Sigma}_{ff}^{-1/2} \boldsymbol{\Sigma}_{fp} \boldsymbol{\Sigma}_{pp}^{-1/2}) \boldsymbol{\nu} \\ \text{s.t.} \ & \boldsymbol{\upsilon}^T \boldsymbol{\upsilon} = 1, \ \ \boldsymbol{\nu}^T \boldsymbol{\nu} = 1 \end{aligned} \tag{17}$$

whose solution is given by the singular value decomposition of the scaled Hankel matrix, $\mathbf{H}$:

$$\mathbf{H} = \boldsymbol{\Sigma}_{ff}^{-1/2} \boldsymbol{\Sigma}_{fp} \boldsymbol{\Sigma}_{pp}^{-1/2} = \boldsymbol{\mathcal{U}} \boldsymbol{\Sigma} \boldsymbol{\mathcal{V}}^T \tag{18}$$

where $\boldsymbol{\mathcal{U}} = \begin{bmatrix} \boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \ldots, \boldsymbol{\upsilon}_r \end{bmatrix}$ and $\boldsymbol{\mathcal{V}} = \begin{bmatrix} \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_r \end{bmatrix}$ are the left and right singular matrices, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ is the diagonal matrix of descending non-zero singular values, and r is the rank of $\mathbf{H}$. The singular values, σ_i, are the maximum solutions of (17), called *canonical correlations*.

Using the projection matrix $\mathbf{J} = \boldsymbol{\mathcal{V}}^T \boldsymbol{\Sigma}_{pp}^{-1/2} \in \mathbb{R}^{r \times \ell p}$, the past data, $\mathbf{Y}_p^+$, can be transformed into the space spanned by maximally correlated *canonical variates*, $\boldsymbol{\mathcal{Z}}_p$, as in

$$\boldsymbol{\mathcal{Z}}_p = \mathbf{J}\mathbf{Y}_p^+ \in \mathbb{R}^{r \times (N-p+1)}. \tag{19}$$

However, based on the values of σ_i, only $n \, (< r)$ dominant singular values truly explain the system dynamics, which corresponds to the first n most correlated canonical variates. Thus, the total space spanned by $\boldsymbol{\mathcal{Z}}_p$ can be partitioned into the *state* subspace $\boldsymbol{\mathcal{Z}}_p^{(S)}$ and *residual* subspace $\boldsymbol{\mathcal{Z}}_p^{(R)}$ [36]. In system identification, only $\boldsymbol{\mathcal{Z}}_p^{(S)}$ is of concern, since it already gives the estimated state sequence, $\widehat{\mathbf{X}}$:

$$\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_k] := \boldsymbol{\mathcal{Z}}_p^{(S)} = \mathbf{J}_n \mathbf{Y}_p^+ \in \mathbb{R}^{n \times (N-p+1)} \tag{20}$$

where $\mathbf{J}_n$ consists of the first n rows of $\mathbf{J}$. The choice of n effectively determines the order of the system. Eq. (20) implies that the states $\widehat{\mathbf{X}}$ in CVA are estimated as linear combinations of a moving time-window of $\mathbf{t}_k$ features.

C. Least-squares Regression

After states estimation, the state-space matrices, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are estimated via multivariate linear least-squares regression [37], [38]. This is the final step in FR-MKCVA.

Using $\mathbf{T}$ from (8) and letting $\mathbf{Y} = [\mathbf{y}_k] \in \mathbb{R}^{m_y \times N}$, the state-space matrices are obtained as:

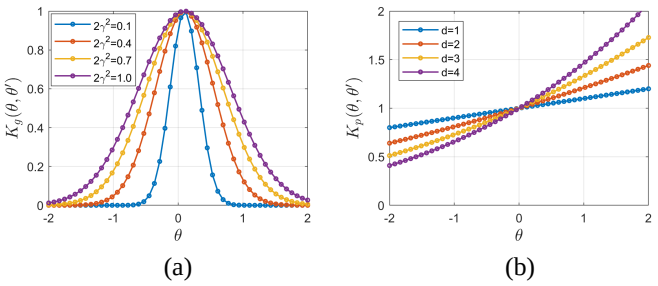
$$[\widehat{\mathbf{C}} \ \widehat{\mathbf{D}}] = \mathbf{Y}_{(:,\,(p+1):N)} \begin{bmatrix} \widehat{\mathbf{X}}_{(:,\,1:(N-p))} \\ \mathbf{T}_{(:,\,p:(N-1))} \end{bmatrix}^{\dagger} \tag{21}$$

$$[\widehat{\mathbf{A}} \ \widehat{\mathbf{B}}] = \widehat{\mathbf{X}}_{(:,\,2:(N-p+1))} \begin{bmatrix} \widehat{\mathbf{X}}_{(:,\,1:(N-p))} \\ \mathbf{T}_{(:,\,p:(N-1))} \end{bmatrix}^{\dagger} \tag{22}$$

where the superscript † represents the pseudo-inverse operation and the subscripts follow MATLAB notation. At this point, the model has now been identified.

Remark 1: Note that the $\mathbf{T}$ regressors were lagged by one time step from the $\mathbf{Y}$ regressors in (21). This is necessary since both $\mathbf{T}$ and $\mathbf{Y}$ contain $\mathbf{y}_k$ information. With lagged regressors, the supposed $\mathbf{t}_k := f(\mathbf{u}_k, \mathbf{y}_k)$ instead becomes $\mathbf{t}_k := f(\mathbf{u}_{k-1}, \mathbf{y}_{k-1})$ in (3)-(4), making the model *causal*. However, during simulation, $\mathbf{t}_k := f(\mathbf{u}_k, \mathbf{y}_{k-1})$ is finally used, assuming that the solved matrices from (21)-(22) remain applicable. This is done so that outputs $\mathbf{y}_k$ can respond to an input $\mathbf{u}_k$ at the same instant k (cf. Fig. 1).

Lastly, since data normalization was done prior to the whole process, the simulated data from the FR-MKCVA model, $\mathbf{y}_k^{\text{sim}}$, must be reverted to their original scale and mean values.



(a) (b)

Fig. 2. Sample plots of kernel values, where $\theta' = 0.1$, for: (a) a local kernel (RBF kernel); and (b) a global kernel (polynomial kernel) [27].

IV. Proposed Kernel Design

Mercer's theorem gives conditions for which functions can act as a dot product in a possibly ∞-dimensional space F [34]. Hence, only those functions that satisfy Mercer's condition can be used as kernels in (5). Common kernel functions [39] include the radial basis function (RBF) kernel:

$$K_g(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2\gamma^2} \right) \tag{23}$$

as well as the polynomial kernel:

$$K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle + 1)^d \tag{24}$$

where γ and d are hyper-parameters, and $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ are training and test samples, respectively. These kernels satisfy Mercer's conditions for $\gamma \neq 0$ and $d \in \mathbb{N}$ [27].

A. Basis for the New Kernel Design

Jordaan [25] had pointed out that the RBF and polynomial kernels are the typical examples of the two main categories of kernels: *local* and *global*, respectively. Fig. 2 plots the behavior of these kernels at varying kernel parameters, with $\theta' = 0.1$ as the training sample. In Fig. 2(a), the local kernel is shown to influence a mapping around the training sample, θ'. However, poor extrapolation is exemplified by the tendency of the mappings to vanish to zero as the input sample, θ, moves away from θ'. Hence, local kernels are known to have good interpolation but poor extrapolation ability [39]. On the other hand, Fig. 2(b) shows that the global kernel has good extrapolation ability, as exemplified by the ability of θ' to influence non-zero mappings in the entire domain of inputs, θ. However, the global kernel has minimal local effect around θ'.

In essense, both classes of kernels have limited interpolation and extrapolation abilities on their own [25]–[27].

By combining the two most commonly used kernels from each class, one can obtain good interpolation and extrapolation abilities simultaneously [25]. In this work, this mixed kernel concept is adopted in the KPCA step of FR-MKCVA. Furthermore, we allow the RBF kernel, K_g, to have different values of γ for each individual feature in $\boldsymbol{\theta}$. By collecting the kernel widths as $\boldsymbol{\Gamma} = \mathrm{diag}(1/(2\gamma_1^2), 1/(2\gamma_2^2), \ldots, 1/(2\gamma_\ell^2))$, the feature-relevant RBF kernel can be written as:

$$K_r(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{\Gamma} (\boldsymbol{\theta} - \boldsymbol{\theta}')\right). \quad (25)$$

Considering a convex mixture of single kernels, the following feature-relevant mixed kernel is proposed:

$$K_{\mathrm{mix}} = \omega K_p + (1 - \omega) K_r \quad (26)$$

where $\omega \in [0, 1]$ is a scalar weight, and K_p and K_r are given in (24) and (25), respectively. Note that the mixed kernel reduces to either single kernel at $\omega = 1$ and $\omega = 0$.

In the K_r kernel, the more relevant features in $\boldsymbol{\theta}$ would have corresponding smaller γ values, and hence, their variation can have more influence in the overall model. By including feature relevance into the kernel, it is postulated that the model would have greater flexibility. However, the relevance of each feature is unknown *a priori*. In addition, the influence of each single kernel, as dictated by ω, is also unknown. In this work, these hyper-parameters, namely $\boldsymbol{\Gamma}$ and ω, together with the CVA model order n, are to be tuned by hold-out validation.

B. Model validation via Random Search

In FR-MKCVA, different values of hyper-parameters give different models. In this paper, these values are tuned by validating the performance of FR-MKCVA on a held out data set that is independent from the training data. Note that the polynomial degree in (26) is set to 1, i.e. a linear kernel, to maximize its role for extrapolation [25]–[27].

In this work, *random search* is used to find optimal values of the hyper-parameters. In random search, the values of each hyper-parameter are uniformly drawn from a predefined range. After drawing a fixed number of times, the settings which produced the best performance on validation data is chosen. The advantages of random search are the following:

1) Recently, Bergstra and Bengio [32] established random search as the baseline strategy for hyper-parameter tuning. They have demonstrated that with the same number of runs, random search is more likely to find the optimal values than grid search or manual search, especially for high-dimensional search spaces.
2) Other population-based methods such as particle swarm optimization or genetic algorithms have a higher computational cost because the objective function needs to be evaluated more times to give the optimum. Hence, random search is more practical and more efficient.

The performance of a model is defined in this work as the normalized mean square error (NMSE):

$$\mathrm{NMSE} = 100\% \times \left(1 - \frac{\|y_{\mathrm{ref}}(t) - y(t)\|^2}{\|y_{\mathrm{ref}}(t) - \mathrm{mean}(y_{\mathrm{ref}}(t))\|^2}\right) \quad (27)$$

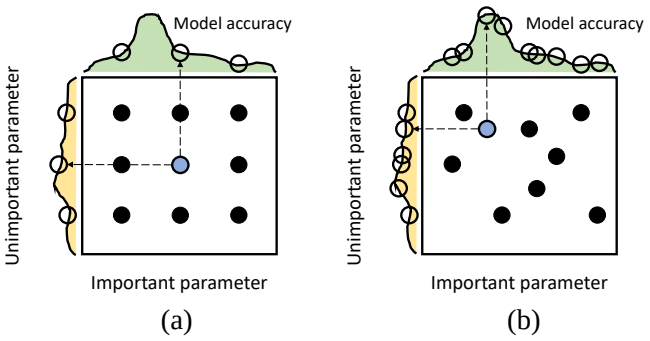


Fig. 3. Comparison between (a) grid search; and (b) random search for hyper-parameter tuning. The nine points denote the candidates. The curves on the left and on the top denote model accuracy (e.g. NMSE) as a function of each search dimension. See more details from [32].

where $y_{\mathrm{ref}}(t)$ are output signals from the actual system, and $y(t)$ are the output signals from the model being tested. In MIMO systems, the NMSE averaged for all output variables is referred to as Mean NMSE for the rest of this paper. In essence, by maximizing the NMSE on a validation data set, random search automatically determines the relevance of each process variable, as well as the mixture weight ω and model order n. Algorithm 1 gives the overall FR-MKCVA method based on the details in Sections III and IV.

Remark 2: Note that the settings for the two KPCAs in lines 3-4 of Algorithm 1 may not necessarily be the same. In this work, if the settings ω and $\boldsymbol{\Gamma}$ are determined for the KPCA on $[\mathbf{u}_k\ \mathbf{y}_k]$, the settings adopted for KPCA on $[\mathbf{y}_k]$ are $\omega' := \omega$ and $\boldsymbol{\Gamma}' := \boldsymbol{\Gamma} \cdot \frac{m}{m_y}$. This scheme is done knowing that γ must depend on the dimensionality of $\boldsymbol{\theta}$ in (23) in order to extract salient features from the data [35], [39].

Remark 3: We prefer to use random search rather than grid search because it was observed in [32] that for learning machines, different hyper-parameters have varying sensitivities to model accuracy. It is not known beforehand which hyper-parameter subspace is more important to explore, and the answer depends on the training data as well. Since grid search explores each search dimension equally, it needlessly explores the less important ones, hence it is less efficient. To illustrate this idea, Fig. 3 shows how the optimal point (blue) found by random search can correspond to a more accurate model than the one found by grid search, with both having the same number of trial points.

V. Case Study

In this section, the performance of the proposed approach is evaluated using an industrial case study. In general, models can be tested on either a simulation task or a multi-step ahead prediction task [3]. Since the former is more stringent, simulation tests are used to demonstrate the advantages offered by the proposed FR-MKCVA.

A. Process Description

The Newell-Lee evaporator system [40], shown in Fig. 4, is a well-known nonlinear case study in process control.

Algorithm 1 Proposed FR-MKCVA with Random Search

Input: $\mathbf{u}_k, \mathbf{y}_k$ training data; $\mathbf{u}'_k, \mathbf{y}'_k$ validation data; I no. of iterations; ranges of $\omega, \mathbf{\Gamma}, n$.
Output: Nonlinear model, Eq. (3)-(4).
Initialisation : $\text{NMSE}_{\min} = \infty$.
for $i = 1$ to I **do**
  Draw a random $\omega_i, \mathbf{\Gamma}_i, n_i$ uniformly.
  Obtain $[\mathbf{t}_k]$ via KPCA on $[\mathbf{u}_k\ \mathbf{y}_k]$, Eq. (26), (5)-(8).
  Obtain $[\mathbf{s}_k]$ via KPCA on $[\mathbf{y}_k]$, Eq. (26), (5)-(8).
  Obtain $[\widehat{\mathbf{x}}_k]$ via CVA, Eq. (10)-(15), (18)-(20).
  Perform least-squares regression, Eq. (21)-(22).
  Simulate the model using $\mathbf{u}'_k$ and compute NMSE_i.
  if $\text{NMSE}_i < \text{NMSE}_{\min}$ **then**
    Assign $\text{NMSE}_{\min} := \text{NMSE}_i$.
    Record $(\omega, \mathbf{\Gamma}, n)_{\text{opt}} := (\omega_i, \mathbf{\Gamma}_i, n_i)$.
    Record $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}, \widehat{\mathbf{D}}, f(\cdot)$.
  end if
end for
return $(\omega, \mathbf{\Gamma}, n)_{\text{opt}}$ and $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}, \widehat{\mathbf{D}}, f(\cdot)$.

The forced-circulation evaporator itself aims to increase the concentration of the feed liquor from X_1 to X_2 by boiling away the solvent using the heat from steam flowing at F_{100}. The removed vapor is condensed by cooling water and exits at a rate of F_5. To stabilize the system, the separator liquid level L_2 is maintained at $L_2^{\text{sp}} = 1.0$ m by manipulating the product flow rate F_2 using a P-controller with $K_c = 5$. The observed variables are $\mathbf{u} = [F_{200},\ P_{100},\ F_3]^T$ and $\mathbf{y} = [L_2,\ P_2,\ X_2]^T$. Simulations of the following 3-state process were carried out in MATLAB Simulink, available in [41], together with added nonlinearities such as saturation limits on variables and valve dynamics (see Table I for nominal values):

$$\frac{dL_2}{dt} = \frac{F_1 - F_4 - F_2}{20} \tag{28}$$

$$\frac{dX_2}{dt} = \frac{F_1 X_1 - F_2 X_2}{20} \tag{29}$$

$$\frac{dP_2}{dt} = \frac{F_4 - F_5}{4} \tag{30}$$

$$\text{where} \begin{cases} T_2 &= 0.5616P_2 + 0.3126X_2 + 48.43 \\ T_3 &= 0.507P_2 + 55.0 \\ F_4 &= \frac{Q_{100} - 0.07F_1(T_2 - T_1)}{38.5} \\ T_{100} &= 0.1538P_{100} + 90.0 \\ Q_{100} &= 0.16(F_1 + F_3)(T_{100} - T_2) \\ F_{100} &= Q_{100}/36.6 \\ Q_{200} &= \frac{0.9576F_{200}(T_3 - T_{200})}{0.14F_{200} + 6.84} \\ T_{201} &= T_{200} + \frac{Q_{200}}{0.07F_{200}} \\ F_5 &= Q_{200}/38.5. \end{cases}$$

Several data sets were produced, each having 1001 samples of the 6 variables at 0.5 min sampling frequency. The data sets differ in the seeds for random noise and input disturbances. The disturbances are APRBS (amplitude-modulated pseudo-random bit signals) [3], as shown in Fig. 5(a). Sample output data sets are also shown in Fig. 5(b).

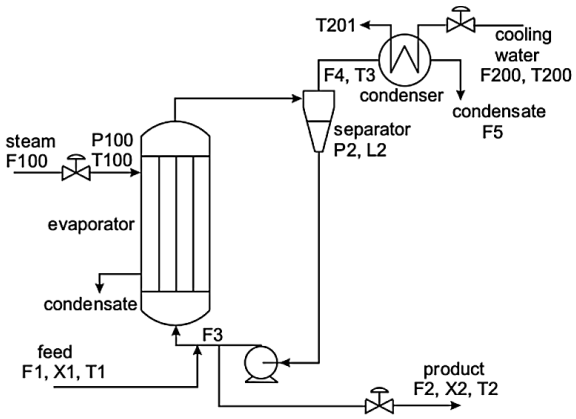


Fig. 4. Process flowsheet of the Newell-Lee evaporator system.

TABLE I
NOMINAL VALUES OF PROCESS VARIABLES

Variable	Description	Value	Unit
F_1	Feed flowrate	10.0	kg/min
F_2	Product flowrate	2.0	kg/min
F_3	Circulating flowrate	50.0	kg/min
F_4	Vapor flowrate	8.0	kg/min
F_5	Condensate flowrate	8.0	kg/min
X_1	Feed composition	5.0	%
X_2	Product composition	25.0	%
T_1	Feed temperature	40.0	°C
T_2	Product temperature	84.6	°C
T_3	Vapor temperature	80.6	°C
L_2	Separator level	1.0	m
P_2	Operating pressure	50.5	kPa
F_{100}	Steam flowrate	9.27	kg/min
T_{100}	Steam temperature	119.9	°C
P_{100}	Steam pressure	194.7	kPa
Q_{100}	Heat duty	339.2	kW
F_{200}	Cooling water flowrate	208.0	kg/min
T_{200}	Inlet CW temperature	25.0	°C
T_{201}	Outlet CW temperature	46.15	°C
Q_{200}	Condenser duty	308.0	kW

Remark 4: Since the case study includes a level controller, closed-loop identifiability is a possible concern. This issue arises when an input-output pair of variables are linked by a feedback loop across a controller, thereby leading to a model that identifies the inverse of this controller rather than the system itself. In this paper, however, none of the input-output pairs are linked in a control loop. Instead, the existing control loop (i.e. the level controller) is treated as part of the system to be identified. One advantage of this treatment is that when the model is used for control design, the interaction from the level control loop will be considered automatically. Hence, closed-loop identification is not an issue in our case study.

B. FR-MKCVA: Training Phase

For the training phase, a single training data set and a single validation data set were generated. To run FR-MKCVA, the number of iterations in the random search was set to $I = 100$. The ranges of the hyper-parameters explored were $\omega \in [0, 1]$, $\gamma_j = 10^a$, $a \in [0, 4]$, and $n \in [3, 10]$. After making two

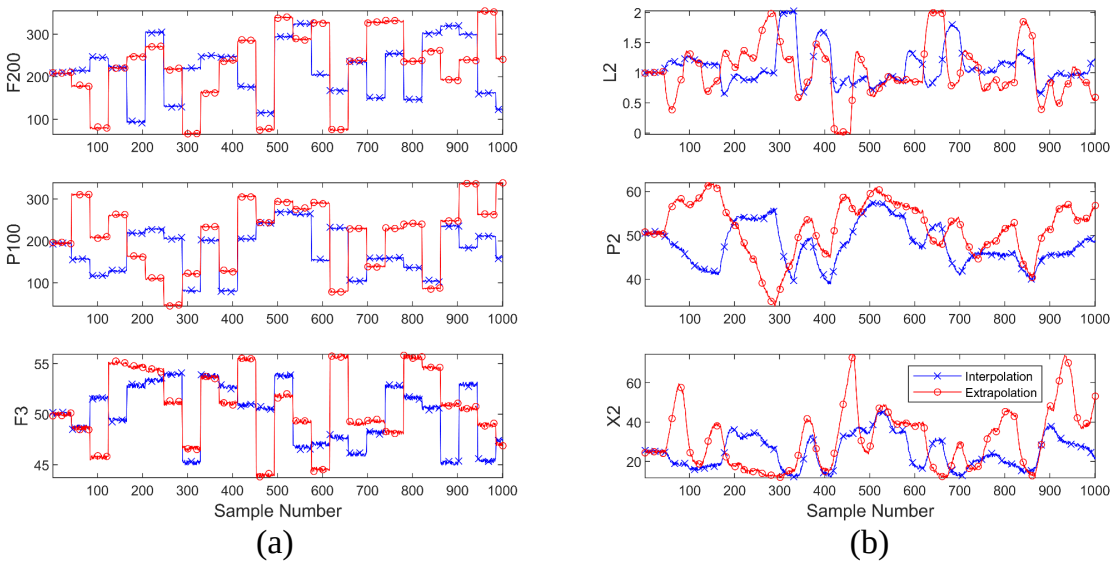


Fig. 5. Sample data sets for: (a) inputs; and, (b) outputs. Blue Xs - sample data used in training, validation, and interpolation tests. Red circles - sample data used in extrapolation tests.

TABLE II
HYPER-PARAMETER VALUES OF MODELS UNDER COMPARISON

Model	Kernel Parameters			Order	Mean NMSE on validation data
	ω	γ	d	n	
FR-MKCVA-1†	0.638	‡	1	4	94.7557 %
FR-MKCVA-2†	0.550	§	1	4	94.3880 %
NARX	-	-	-	-	94.1646 %
MKCVA-1†	0.810	6.948	1	4	94.1125 %
MKCVA-2†	0.609	9.083	1	4	93.9209 %
KCVA-RBF†	-	59.459	-	4	93.8312 %
KCVA-POLY†	-	-	1	5	89.5289 %
CVA-1	-	-	-	8	88.0721 %
CVA-2	-	-	-	4	84.7931 %

† Values were found by maximizing the Mean NMSE on validation data using random search. See (26) for the meanings of ω, γ, d.
‡ [14.01, 603.49, 191.04, 1.79, 7924.30, 2.72].
§ [1341.92, 8.90, 132.89, 2.32, 41.81, 2.63].

separate runs of Algorithm 1, Table II reports two different FR-MKCVA models, namely FR-MKCVA-1 and 2. Although the settings of these models are different, they resulted in similar Mean NMSEs of 94%. This consistency makes random search a valid way to tune the proposed FR-MKCVA models.

For comparison, Table II also lists 5 other nonlinear models trained on the same training data and tuned by the same validation data via random search: 1 closed-loop NARX, 2 MKCVA models, and 2 single-kernel CVA models. The closed-loop Nonlinear AutoRegressive network with eXogenous inputs (NARX) is a single hidden layer neural net whose structure includes output feedback. The output feedback aspect is also present in (3)-(4) since $\mathbf{t}_k$ is a function of $\mathbf{y}_{k-1}$. The NARX was trained with Bayesian regularization, while random search tuned the structure to have 10 hidden neurons, as chosen from the range $[2, 15]$. The two MKCVA models were trained in the same way as FR-MKCVA except that the RBF kernel was specified by only a single γ rather than $\mathbf{\Gamma}$. Meanwhile, KCVA-RBF and KCVA-POLY only used a single RBF and a single polynomial kernel, respectively. When these KCVA models were tuned, the same range of exploration of hyper-parameters as those in the FR-MKCVA were used. In KCVA-POLY, the polynomial degree was additionally explored in the range $d \in [1, 5]$. Finally, two CVA models are to be compared as well. The order of CVA-1 was obtained from the knee of the singular values plot, which suggests the threshold between the state space and residual space [36]. Meanwhile, CVA-2 has a model order that was matched to that in FR-MKCVA.

As seen on Table II, the FR-MKCVA models achieved the highest Mean NMSEs on the same validation data. Also, the tuned γ values of L_2 and X_2 were consistently small between the two models. This means that model validation has deemed their relevance to be large in order to achieve good performance. But despite having high NMSEs, the superiority of FR-MKCVA must not be concluded from these results. All the models must be evaluated under multiple test data sets that require both interpolation and extrapolation.

C. FR-MKCVA: Testing Phase

For the testing phase, 50 test data sets were generated each for *interpolation* and *extrapolation tests*. The disturbance amplitude used in the interpolation data sets is the same as that in the training and validation data sets. On the other hand, the extrapolation data sets were produced with 25% larger amplitude. Sample data sets are presented in Fig. 5.

Interpolation test results are summarized in Fig. 6(a). The figure shows that the FR-MKCVA models perform the best among the models being compared, having 94% median NMSEs. Next to these, the NARX, MKCVA, and KCVA-RBF models are shown to have similar accuracies, with median NMSEs of 90%. High accuracies can be attributed to the flexibility of the RBF kernel and neural nets in modelling nonlinearities. More importantly, results show that adding feature relevance to the kernel design clearly improves the accuracy

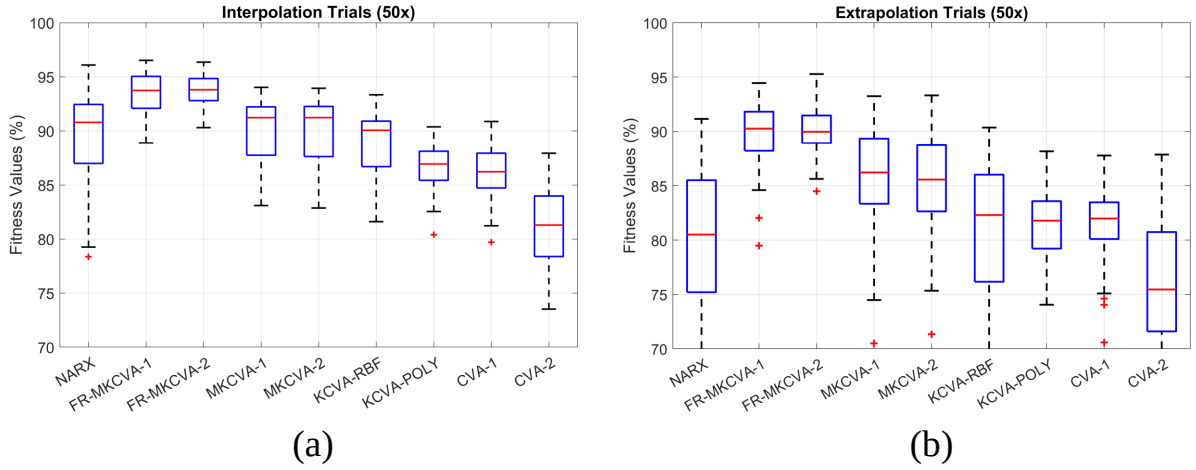


Fig. 6. Simulation performance for models under comparison: (a) interpolation tests; and (b) extrapolation tests. See Table II for model details. The boxplots summarize the Mean NMSE (fitness) results on 50 test data sets that differ in input sequences and random noise.

of mixed kernel CVA models. In contrast, the remaining CVA models have poor accuracy due to being only linear models. Note that the nonlinear models incurred a large improvement against CVA-2, even if they have the same order of $n = 4$. Hence, the use of nonlinear models at all is justified for the system identification of the evaporator case study.

Extrapolation test results are then summarized in Fig. 6(b). First, we demonstrate the advantage of mixed kernel designs. Note that in Fig. 6(b), the single-kernel CVA models now have similar median NMSEs with the linear CVA models during extrapolation. For instance, the KCVA-RBF model produced even lower NMSEs than CVA-1 in some of the trials, and the same is true with the NARX. This means that the RBF kernel and the NARX are poor extrapolators. In contrast, the MKCVA models retained their superiority to the linear CVA models due to the combined advantages from the RBF (local) and polynomial (global) kernels. This benefit is also present in the FR-MKCVA models. Hence, mixed kernel designs have better generalization ability than single-kernel ones. Second, Fig. 6(b) also shows that the FR-MKCVA models remained superior to the rest of the models owing to the feature relevance idea. For extrapolation, FR-MKCVA models incurred median NMSEs of 90% , compared to only 85% for the MKCVA models. This means that allowing different kernel widths in the RBF kernel portion of the mixed kernel can increase the model flexibility. When these parameters are tuned properly, the model can generalize better from the training data.

To further appreciate the performance of FR-MKCVA models, Fig. 7 shows scatter comparison plots of the models FR-MKCVA-1, NARX, KCVA-RBF, and CVA, when used to simulate the X_2 variable in one interpolation and one extrapolation data set. X_2 is an important variable in the process as it indicates the quality of the desired product from the evaporator. Fig. 7(a) mainly shows that the X_2 values predicted by the FR-MKCVA model are the closest to the actual observed values. The predictions from other models were less accurate, especially at too low or too high values of X_2 (see linear CVA and the NARX). This means that these models did not capture the system behavior well at these

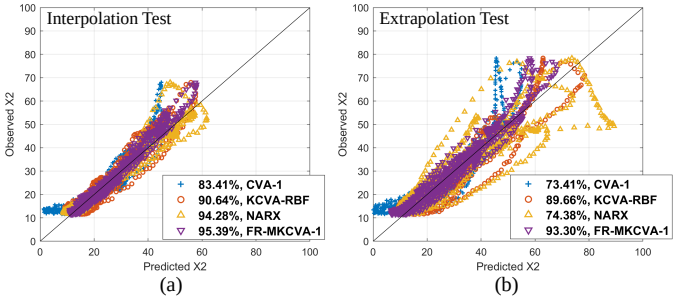


Fig. 7. Performance comparison of FR-MKCVA against NARX, KCVA-RBF, and CVA in a sample X_2 data set; (a) interpolation test; (b) extrapolation test. See Table II for model details. NMSE fitness values are also reported for each method.

regions. For the extrapolation test in Fig. 7(b), it can be seen that both KCVA-RBF and the NARX models incurred large decreases in accuracy. In particular, the behavior of the NARX became too erratic to be a reliable predictor of the X_2. It can be said that these models have poor extrapolation abilities. In contrast, the FR-MKCVA predictions remained close to the observed data in Fig. 7(b). Hence, the ideas of using mixed kernels and feature-relevance are valid improvements to kernel design for system identification.

We further note that the performance of the NARX can be improved in these experiments by training it with more data than just 1001 samples. The kernel-based SIMs, on the other hand, may become too slow for prediction when trained with larger data sets [42]. Nevertheless, our proposed FR-MKCVA is intended for scenarios where training data has a moderate sample size but is high-dimensional. For these scenarios, we have shown that certain improvements in the kernel design can produce better kernel-based SIMs. This calls for more research into kernel design in the future, such as exploring deep and ensemble kernel architectures [42]. By improving model generalization ability, industrial activities that depend on these models can perform better.

VI. Conclusion

In this paper, we propose a new nonlinear subspace identification method, called Feature-Relevant Mixed Kernel Canonical Variate Analysis (FR-MKCVA), for MIMO processes. The main contribution of this study is to improve the kernel design of kernel CVA in two ways: (1) by combining the commonly used single kernels, namely the radial basis function (RBF) and polynomial kernel into a single model; and, (2) by allowing the RBF kernel width to vary in each dimension to account for the relevance of the process variables to the overall model. This paper also presents a methodology for training the model via kernel PCA followed by CVA and linear regression. Model validation using random search was also found to be an effective and practical method for tuning the model hyperparameters.

An industrial evaporator case study was used to demonstrate the effectiveness of the proposed identification method. Results show the superiority of FR-MKCVA to single-kernel CVA, linear CVA, and the NARX neural network models under both interpolation and extrapolation scenarios. In the future, more research into deep or ensemble kernel designs can be ventured towards increasing model flexibility. However, this must be coupled with developments in model validation and hyperparameter tuning in order to produce accurate models in a fast and practical way.

References

[1] L. Ljung, "Perspectives on system identification," *Annu. Rev. Control*, vol. 34, DOI 10.1016/j.arcontrol.2009.12.001, no. 1, pp. 1–12, 2010.

[2] X. Hong, R. J. Mitchell, S. Chen, C. J. Harris, K. Li, and G. W. Irwin, "Model selection approaches for non-linear system identification: A review," *Int. J. Syst. Sci.*, vol. 39, DOI 10.1080/00207720802083018, no. 10, pp. 925–946, 2008.

[3] O. Nelles, *Nonlinear System Identification*. Springer-Verlag, 2001.

[4] L. Ljung, *System Identification: Theory for User*. Prentice Hall, 1987.

[5] V. Verdult, J. Suykens, J. Boets, I. Goethals, and B. De Moor, "Least squares support vector machines for kernel CCA in nonlinear state-space identification," *Proc. 16th Int. Symp. Math. Theory Networks Syst. (MTNS 2004)*, pp. 1–11, 2004.

[6] S. J. Qin, "An overview of subspace identification," *Comput. Chem. Eng.*, vol. 30, DOI 10.1016/j.compchemeng.2006.05.045, no. 10-12, pp. 1502–1513, Sep. 2006.

[7] T. Katayama, *Subspace Methods for System Identification*. Springer London, 2005.

[8] A. Simoglou, E. B. Martin, and A. J. Morris, "Statistical performance monitoring of dynamic multivariate processes using state space modelling," *Comput. Chem. Eng.*, vol. 26, DOI 10.1016/S0098-1354(02)00012-1, no. 6, pp. 909–920, 2002.

[9] L. Ljung, "Aspects and Experiences of User Choices in Subspace Identification Methods," *IFAC Proc. Vol.*, vol. 36, DOI 10.1016/S1474-6670(17)35015-2, no. 16, pp. 1765–1770, Sep. 2003.

[10] C. Schaper, W. Larimore, D. Seborg, and D. Mellichamp, "Identification of chemical processes using canonical variate analysis," *Comput. Chem. Eng.*, vol. 18, no. 1, pp. 55–69, 1994.

[11] J. Lu and F. Liu, "Statistical Modeling of Dynamic Multivariate Process Using Canonical Variate Analysis," in *2006 Int. Conf. Inf. Autom.*, DOI 10.1109/ICINFA.2006.374115, pp. 218–221. IEEE, Dec. 2006.

[12] B. C. Juricek, D. E. Seborg, and W. E. Larimore, "Identification of the Tennessee Eastman challenge process with subspace methods," *Control Eng. Pract.*, vol. 9, DOI 10.1016/S0967-0661(01)00124-1, no. 12, pp. 1337–1351, 2001.

[13] C. Ruiz-Cárcel, L. Lao, Y. Cao, and D. Mba, "Canonical variate analysis for performance degradation under faulty conditions," *Control Eng. Pract.*, vol. 54, DOI 10.1016/j.conengprac.2016.05.018, pp. 70–80, 2016.

[14] V. Verdult, "Nonlinear System Identification: A State-Space Approach," Ph.D. dissertation, University of Twente, 2002.

[15] A. Garulli, S. Paoletti, and A. Vicino, "A survey on switched and piecewise affine system identification," *IFAC Proc. Vol.*, vol. 45, DOI 10.3182/20120711-3-BE-2027.00332, no. 16, pp. 344–355, Jul. 2012.

[16] R. Al Seyab and Y. Cao, "Differential recurrent neural network based predictive control," *Comput. Chem. Eng.*, vol. 32, DOI 10.1016/j.compchemeng.2007.07.007, no. 7, pp. 1533–1545, 2008.

[17] R. Kumar, S. Srivastava, J. R. P. Gupta, and A. Mohindru, "Comparative study of neural networks for dynamic nonlinear systems identification," *Soft Comput.*, DOI 10.1007/s00500-018-3235-5, May. 2018.

[18] R. Babuška and H. Verbruggen, "Neuro-fuzzy methods for nonlinear system identification," *Annu. Rev. Control*, vol. 27 I, DOI 10.1016/S1367-5788(03)00009-9, pp. 73–85, 2003.

[19] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, DOI 10.1016/j.automatica.2014.01.001, no. 3, pp. 657–682, 2014.

[20] R. T. Samuel and Y. Cao, "Kernel canonical variate analysis for nonlinear dynamic process monitoring," *IFAC-PapersOnLine*, vol. 28, DOI 10.1016/j.ifacol.2015.09.034, no. 8, pp. 605–610, 2015.

[21] S. Van Vaerenbergh, J. Vía, and I. Santamaría, "Adaptive Kernel Canonical Correlation Analysis Algorithms for Nonparametric Identification of Wiener and Hammerstein Systems," *EURASIP J. Adv. Signal Process.*, DOI 10.1155/2008/875351, no. 1, p. 875351, 2008.

[22] P. L. Lai and C. Fyfe, "Kernel and Nonlinear Canonical Correlation Analysis," *Int. J. Neural Syst.*, vol. 10, DOI 10.1142/S012906570000034X, no. 5, pp. 365–377, Oct. 2000.

[23] Y. Kawahara, T. Yairi, and K. Machida, "A Kernel Subspace Method by Stochastic Realization for Learning Nonlinear Dynamical Systems," *Neural Inf. Process. Syst.*, vol. 19, pp. 665–672, 2007.

[24] I. Goethals, L. Hoegaerts, V. Verdult, J. A. K. Suykens, and B. De Moor, "Subspace Identification of Hammerstein-Wiener systems using Kernel Canonical Correlation Analysis," 2004.

[25] E. M. Jordaan, "Development of Robust Inferential Sensors: Industrial Applications of Support Vector Machines for Regression," Ph.D. dissertation, Technische Universiteit Eindhoven, 2002.

[26] X. Zhu, Z. Huang, H. Tao Shen, J. Cheng, and C. Xu, "Dimensionality reduction by mixed kernel canonical correlation analysis," *Pattern Recognit.*, vol. 45, DOI 10.1016/j.patcog.2012.02.007, no. 8, pp. 3003–3016, 2012.

[27] K. E. S. Pilario, Y. Cao, and M. Shafiee, "Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes," *Comput. Chem. Eng.*, vol. 123, DOI 10.1016/j.compchemeng.2018.12.027, pp. 143–154, Apr. 2019.

[28] X. Zhao and Y. Xue, "Output-relevant fault detection and identification of chemical process based on hybrid kernel T-PLS," *Can. J. Chem. Eng.*, vol. 92, DOI 10.1002/cjce.22031, no. 10, pp. 1822–1828, 2014.

[29] Y. Chen, M. Kloft, Y. Yang, C. Li, and L. Li, "Mixed kernel based extreme learning machine for electric load forecasting," *Neurocomputing*, vol. 312, DOI 10.1016/j.neucom.2018.05.068, pp. 90–106, 2018.

[30] X.-Y. Wang and M. Han, "Multivariate time series prediction based on multiple kernel extreme learning machine," *2014 Int. Jt. Conf. Neural Networks*, vol. 61, DOI 10.1109/IJCNN.2014.6889479, no. 8, pp. 198–201, 2014.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[32] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[33] R. Castro-Garcia, O. M. Agudelo, K. Tiels, and J. A. Suykens, "Hammerstein system identification using LS-SVM and steady state time response," *2016 Eur. Control Conf.*, DOI 10.1109/ECC.2016.7810430, pp. 1063–1068, 2017.

[34] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[35] S. W. Choi, C. Lee, J. M. Lee, J. H. Park, and I. B. Lee, "Fault detection and identification of nonlinear processes based on kernel PCA," *Chemom. Intell. Lab. Syst.*, vol. 75, DOI 10.1016/j.chemolab.2004.05.001, no. 1, pp. 55–67, 2005.

[36] P.-E. Odiowei and Y. Cao, "Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations," *IEEE Trans. Ind. Informatics*, vol. 6, DOI 10.1109/TII.2009.2032654, no. 1, pp. 36–45, 2010.

[37] L. Shang, J. Liu, K. Turksoy, Q. Min Shao, and A. Cinar, "Stable recursive canonical variate state space modeling for time-varying processes," *Control Eng. Pract.*, vol. 36, DOI 10.1016/j.conengprac.2014.12.006, pp. 113–119, 2015.

[38] W. E. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control," *Proc. IEEE Conf. Decis. Control*, vol. 2, DOI 10.1109/CDC.1990.203665, pp. 596–604, 1990.
[39] Y. Bengio, O. Delalleau, and N. L. Roux, "The Curse of Highly Variable Functions for Local Kernel Machines," in *Neural Inf. Process. Syst.*, pp. 107–114, 2006.
[40] R. B. Newell and P. L. Lee, *Applied Process Control - A Case Study*. NJ: Prentice Hall, 1989.
[41] K. E. S. Pilario, "Newell-Lee Evaporator System for System Identification," 2020. [Online]. Available: https://uk.mathworks.com/matlabcentral/fileexchange/68641-newell-lee-evaporator-system-for-system-identification
[42] K. E. Pilario, M. Shafiee, Y. Cao, L. Lao, and S.-h. Yang, "A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring," *Processes*, vol. 8, DOI 10.3390/pr8010024, p. 24, 2020.

Karl Ezra S. Pilario received the B.Sc. *(summa cum laude)* and M.Sc. degrees in Chemical Engineering both from the University of the Philippines, Diliman, Quezon City, Philippines, in 2012 and 2015, respectively. He obtained the Ph.D. degree in Energy and Power from Cranfield University, Cranfield, U.K. in 2020.

He is an Assistant Professor on study leave from the Department of Chemical Engineering, University of the Philippines - Diliman. He also held the Dominador I. Ilio Engineering Centennial Professorial Chair from 2015 to 2017. His current research interests include industrial applications of data analytics and machine learning, including process monitoring.

Dr. Pilario was a recipient of the Limcaoco Young Instructor Award for teaching excellence, undergraduate advising, and mentoring in academic competitions in 2015.

Yi Cao received the M.Sc. degree in Industrial Automation from Zhejiang University, China in 1985 and the Ph.D. degree in Engineering from the University of Exeter, U.K. in 1996.

He is a Professor in the College of Chemical and Biological Engineering, Zhejiang University, China, since 2018. He was a Reader in the School of Water, Energy and Environment, Cranfield University, from 2000 to 2018. He also held several Research Associate positions from the University of Leicester, Loughborough University and the University of Exeter, between 1996 and 2000. He was a post-doctoral researcher in the Department of Chemical Engineering, the Norwegian University of Science and Technology, in 1995 and a Lecturer in the Department of Electrical Engineering, Zhejiang University, between 1985 and 1992. His research interests are in advanced process control, including self-optimizing control, nonlinear system identification, nonlinear model predictive control and process monitoring.

Prof. Cao received the Control Engineering Practice Paper Prize at the Control '96, the Energy Innovation Award from the East England Energy Group in 2010, the best Oral Presentation Award at the ADCHEM 2015.

Mahmood Shafiee (PhD, CEng, MIET, MIMechE, MIAM, MIMA, FHEA, PGCHE) is the Head of Mechanical Engineering Department, a Programme Chair and Reader at School of Engineering and Digital Arts, University of Kent, UK. He is also a Visiting Fellow at Cranfield University as well as a Course Director at the University of Sheffield.

Dr. Shafiee has more than fifteen years' experience of research and consultancy in different fields of Mechanical Systems Engineering, including Reliability-based Design, Autonomous and Robotic Maintenance, Failure of Materials and Structures, Mechanical Degradation and Life Assessment, Structural Health Monitoring (SHM) and Condition Monitoring (CM). He has a very strong track record of developing research proposals and attracting external funding from a variety of funders to undertake research and knowledge exchange projects. To date, he has authored 1 book and 7 invited book chapters, received 3 patents and published over 150 journal papers and conference articles. He has supervised several PhD students, MEng/BEng group projects, and MEng/BEng individual thesis projects through to completion. He has been a keynote speaker and a member of technical committee in over 50 national and international conferences. Dr. Shafiee sits on the European Safety and Reliability Association (ESRA) Committee, and serves as a Member of the Institution of Engineering and Technology (IET), Institution of Mechanical Engineers (IMechE), Institute of Asset Management (IAM), RenewableUK and EERA JP Wind Committees.