Pre-classification based stochastic reduced-order model for time-dependent complex system

Meixin Xiong[1], Liuhong Chen[1], Ju Ming[1*] and Zhiwen Zhang[2]

[1*]School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China.
[2]Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China.

*Corresponding author(s). E-mail(s): jming@hust.edu.cn;
Contributing authors: xmx2018@hust.edu.cn;
liuhong_c@hust.edu.cn; zhangzw@hku.hk;

Abstract

We propose a novel stochastic reduced-order model (SROM) for complex systems by combining clustering and classification strategies. Specifically, the distance and centroid of centroidal Voronoi tessellation (CVT) are redefined according to the optimality of proper orthogonal decomposition (POD), thereby obtaining a time-dependent generalized CVT, and each class can generate a set of cluster-based POD (CPOD) basis functions. To learn the classification mechanism of random input, the naive Bayes pre-classifier and clustering results are applied. Then for a new input, the set of CPOD basis functions associated with the predicted label is used to reduce the corresponding model. Rigorous error analysis is shown, and a discussion in stochastic Navier-Stokes equation is given to provide a context for the application of this model. Numerical experiments verify that the accuracy of our SROM is improved compared with the standard POD method.

Keywords: naive Bayes pre-classifier, generalized centroidal Voronoi tessellation, proper orthogonal decomposition, stochastic reduced-order model, time-dependent problems.

MSC Classification: 60H15 , 62H30 , 60H35 , 76D05

1 Introduction

The reduced-order model (ROM) [1] plays a vital role in large-scale simulations, real-time calculations, and optimal control problems, which first introduces a low-dimensional subspace of the state space, and then calculates the coordinates of the system state in this subspace through projection techniques, also known as reduced state vector. It ensures the essential characteristics of the system while achieves the goal of reducing computational complexity. There are a variety of ways to construct the low-fidelity ROM. Among them, proper orthogonal decomposition (POD) based on the optimal Galerkin projection distance is one of the most successful methods, which has been widely applied in numerous fields, including signal analysis and pattern recognition [2, 3], image processing [4], geophysical fluid dynamics [5–7], and biomedical engineering [8].

In many practical problems, the collected data belongs to categorical data, such as countable qualitative data or grouped quantitative data. Then, the natures of these problems can be further explored through the categorical data analysis [9–12]. Clustering [13, 14] and classification [15] are two advanced tools. Clustering is a method for statistical analysis of data and has become an important part of machine learning. It is a process of dividing a given data set into several subsets according to some defined distances. Its purpose is to maximize the intra-cluster similarity and minimize the inter-cluster similarity under the given distance measure. On the one hand, clustering itself is a statistical analysis technique. On the other hand, it is often used as a tool for data exploration, data cleaning, and data organizing in the pre-process stage of other data analysis methods. In the past few decades, clustering approaches have been applied to the numerical simulations of partial differential equations (PDEs), and one of the most popular methods is centroidal Voronoi tessellation (CVT) [16]. Some of the notable works in this area are as follows: Burkardt et al. in [17] introduced a reduced-order modeling methodology based on CVT for complex systems and in [18] compared the performance of ROMs based on POD and CVT, Du et al. in [19] proposed a hybrid method named CVT based POD for model reduction, and Kaiser et al. in [20] combined the cluster analysis and transition matrix models to propose a novel cluster-based reduced-order modelling strategy for unsteady flows. We refer to [21, 22] for further discussions.

Classification is another method of data statistical analysis, which assigns labels to samples according to their features. This method belongs to supervised learning and includes two parts: classifier learning and the prediction/-classification of new samples. When a new sample is assigned to the class with the highest similarity, using the data in this class to study the sample can make full use of the existing information and eliminate the redundant information brought by the data in other classes. Recently, the ideas of classification have been applied to the study of PDEs. Bright et al. in [23] combined classification and compressive sensing to determine the flow characteristics around a cylinder

and in [24] proposed sparse measurements to classify and reconstruct time-dependent data, and Brunton et al. in [25] developed a classification scheme to determine the region to which the nonlinear dynamical system belongs. More discussions can consult the literatures [26–28]. For a stochastic system, there may be large differences between the realizations of its state in some cases. In order to reduce the model and reconstruct the state better, clustering and classification methods can be combined. The former is used to organize the given data according to similarity, while the latter trains a classifier based on the clustering results for assigning labels to new samples. Then the samples can be studied by using the predicted subsets instead of the entire data set.

In this work, we combine clustering and classification methods to propose a pre-classification based stochastic ROM (SROM) for improving the accuracy of the POD reduced-order solutions of stochastic evolution problems. The method mainly consists of two parts. In the first part, several groups of cluster-based POD (CPOD) basis functions are generated by constructing a time-dependent clustering method. Due to the generalizability of the distance in CVT method, the spatio-temporal projection distance from a function to a multidimensional space is used to define the time-dependent generalized Voronoi tessellation (t-gVT). The corresponding generalized centroid is defined as the subspace spanned by the POD basis functions according to the optimality of POD method. Similar to CVT, the time-dependent generalized CVT (t-gCVT) can be obtained when the generators coincide with the generalized centroids. In order to simplify the construction of t-gCVT, the modified version is introduced by using the time-sampling snapshots to approximately calculate the time integral in generalized distance and using the Monte Carlo (MC) method to estimate the expectation of projection distance. From this, the spatio-temporal data is divided into several classes, and each class can generate a set of snapshot-based POD basis functions. In the second part, we construct the pre-classification based SROM. Considering the mapping relationship between the input and output of the system, we first use the clustering results to train a pre-classifier to provide predicted labels for the new inputs, and then use the CPOD basis functions associated with the labels to reduce their models. Here, the naive Bayes classifier[29, 30] based on the principle of maximum posterior probability is adopted to establish the classification mechanism. We would like to point out that other classifiers, such as k-nearest neighbor[31], decision trees[32], support vector machine[33], etc, can also be combined with our CPOD basis functions without any difficulty. The main ideas of our method are shown in Figures 1 and 2. We call the method of combining CPOD basis functions and naive Bayes pre-classifier to construct SROM as the *CPOD-NB method*.

The rest of this paper is organized as follows. In section 2, we briefly introduce the traditional POD and CVT methods. In section 3, we describe in detail the modified t-gCVT for generating the CPOD basis functions and the naive Bayes method for pre-classification, then combine them to propose the CPOD-NB method for model reduction. The error estimation of the SROM based

on CPOD-NB method and the strategy used for estimating the error rate of naive Bayes pre-classifier are given in section 4. The stochastic Navier-Stokes equation we use as study background is presented in section 5. Numerical experiments are shown in section 6. Finally, some conclusions are given in section 7.

2 Preliminary

We begin by some function spaces and notations needed, then briefly recall the classical POD and CVT methods related to this work.

Denote the system of stochastic partial differential equations (SPDEs) of unknown function u as

$$F(u(t,\mathbf{x},\boldsymbol{\xi});\ \boldsymbol{\xi}) = 0 \qquad (t,\mathbf{x},\boldsymbol{\xi}) \in [0,T]\times D\times\Gamma, \tag{2.1}$$

where function u has proper initial and boundary value conditions, $\mathbf{x}$ is the spatial variable, t is the time variable and $\boldsymbol{\xi}$ could be other parameters. Let $L^2(D)$ be the set of square-integrable functions defined on domain D with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_{L^2(D)}$. We denote the space of all measurable functions $u:[0,T]\to L^2(D)$ by

$$\mathcal{L}^2([0,T];\ L^2(D)) := \left\{u:[0,T]\to L^2(D)\mid u \text{ measurable},\ \|u\|_{\mathcal{L}^2([0,T];\ L^2(D))}<\infty\right\} \tag{2.2}$$

where

$$\|u\|_{\mathcal{L}^2([0,T];\ L^2(D))} = \left(\int_0^T \|u\|^2_{L^2(D)}dt\right)^{1/2}. \tag{2.3}$$

2.1 Proper orthogonal decomposition

Given a positive integer d, for the system of SPDEs (2.1), the POD procedure is to find the orthonormal basis functions $\{\phi_j(\mathbf{x})\}_{j=1}^d$ that minimize error measure

$$\mathcal{E}^{\mathrm{POD}}(\Pi^d) = \mathbb{E}\left[\left\|u-\Pi^d u\right\|^2_{\mathcal{L}^2([0,T];\ L^2(D))}\right], \tag{2.4}$$

where $\mathbb{E}[\cdot]$ denotes expectation, Π^d is a d-dimensional subspace spanned by $\{\phi_j\}_{j=1}^d$, and $\Pi^d u = \sum_{j=1}^d \langle u,\phi_j\rangle\phi_j$ represents the projection of u onto the subspace. By the Lagrange multiplier method, the minimization problem is equivalent to

$$\mathbb{E}\left[\int_0^T \langle u,\phi_j\rangle\, u dt\right] = \lambda_j\phi_j \qquad j=1,2,\ldots,d, \tag{2.5}$$

where $(\{\lambda_j\},\{\phi_j\})$ is called the eigenpair of operator $\mathbf{C}$ defined as

$$\mathbf{C}\phi_j = \mathbb{E}\left[\int_0^T \langle u,\phi_j\rangle\, u dt\right]. \tag{2.6}$$

We use the MC method to estimate the expectation and time-sampling snapshots to calculate the time integral, then (2.6) can be approximated as

$$\mathbf{C}\phi_j = \frac{\Delta t}{n}\sum_{i=1}^{n}\sum_{k=1}^{J}\langle u(t_k,\mathbf{x},\boldsymbol{\xi}_i),\phi_j\rangle\, u(t_k,\mathbf{x},\boldsymbol{\xi}_i), \tag{2.7}$$

where $t_0 = 0$, $\Delta t = T/J$ and $t_k = t_{k-1} + \Delta t$ for $k = 1, 2, \ldots, J$. Denote the snapshot set as

$$\begin{aligned}\mathcal{W} =& [v_1, \ldots, v_{nJ}]\\ :=& [u(t_1,\mathbf{x},\boldsymbol{\xi}_1), \ldots, u(t_J,\mathbf{x},\boldsymbol{\xi}_1), \ldots, u(t_1,\mathbf{x},\boldsymbol{\xi}_n), \ldots, u(t_J,\mathbf{x},\boldsymbol{\xi}_n)].\end{aligned} \tag{2.8}$$

Combining (2.5) and (2.7), the orthonormal POD basis functions can be represented as

$$\phi_j = \frac{1}{\sqrt{nJ\sigma_j}}\sum_{i=1}^{nJ} y_i^{(j)} v_i \qquad j = 1, 2, \ldots, d. \tag{2.9}$$

Here, $\{y_i^{(j)}\}$ and $\{\sigma_j\}$ satisfy the following eigenvalue problem

$$RY = Y\Lambda, \tag{2.10}$$

where the components of matrices R and Y are defined as $R_{ij} = \frac{1}{nJ}\langle v_i, v_j\rangle$ and $Y_{ij} = y_i^{(j)}$ respectively, and $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_{nJ})$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{nJ} \geq 0$ and $\sigma_j = \frac{\lambda_j}{T}$ for $j = 1, 2, \ldots, nJ$. Therefore, with snapshot set (2.8) and POD basis functions (2.9), the minimum value of measure (2.4) can be approximated as

$$\mathcal{E}^{\mathrm{POD}}(\Pi^d) = \sum_{j=d+1}^{nJ}\lambda_j = T\sum_{j=d+1}^{nJ}\sigma_j, \tag{2.11}$$

which is referred to as the "POD energy".

We summarize the above discussions by giving the following MC-based method for generating POD basis functions (Algorithm 1).

2.2 Centroidal Voronoi tessellation

Given a set of functions $U = \{v_i \in L^2(D)\}_{i=1}^n$, the CVT of set U is a special Voronoi tessellation with the centroids $\{z_k^* \in L^2(D)\}_{k=1}^K$ of Voronoi regions

$$\mathcal{U}_k = \{v \in U \mid \mathcal{D}(v, z_k) \leq \mathcal{D}(v, z_i) \text{ for all } i \neq k\} \quad k = 1, 2, \ldots, K \tag{2.12}$$

satisfying $z_k^* = z_k$ for $k = 1, 2, \ldots, K$, where $\{z_k \in L^2(D)\}_{k=1}^K$ are called the generators of set $\{\mathcal{U}_k\}_{k=1}^K$, K refers to the number of clusters, and the

Algorithm 1 MC-based POD method

Input: input set $X = \{\boldsymbol{\xi}_i\}_{i=1}^{n}$, time step Δt, a positive integer d.
Output: POD basis functions $\{\phi_j\}_{j=1}^{d}$.

1: Solve system (2.1) with inputs X and step size Δt to obtain set $\{u(t_k, \mathbf{x}, \boldsymbol{\xi}_i)\}_{i=1,k=1}^{n,\ J}$, which forms the snapshot set $\mathcal{W}$ defined in (2.8).
2: Generate the POD basis functions $\{\phi_j\}_{j=1}^{d}$ defined in (2.9) by solving eigenvalue problem (2.10).

distance can be selected as any metric, for example the $L^2(D)$ distance as $\mathcal{D}(v, z_k) = \|v - z_k\|_{L^2(D)}$ [16]. When the distances between a point v and two generators z_i, z_j are same and the smallest, the principle of random assignment between these two classes is adopted. According to the partition rule of CVT, we know that it minimizes the error measure

$$\mathcal{E}^{\mathrm{CVT}}\left(\{\mathcal{U}_k\}_{k=1}^{K};\ \{z_k\}_{k=1}^{K}\right) = \sum_{k=1}^{K} \sum_{v \in \mathcal{U}_k} \mathcal{D}^2(v, z_k), \tag{2.13}$$

and (2.13) is referred to as the "CVT energy".

There are several methods that can be used to construct a CVT for a given data set. Among them, the iterative-based Lloyd's method (Algorithm 2) [34] is one of the most popular and simplest approaches.

Algorithm 2 Lloyd's method

Input: data set U, a positive integer K, a set of initial generators $\{z_k\}_{k=1}^{K}$.
Output: CVT $\left(\{\mathcal{U}_k\}_{k=1}^{K};\ \{z_k^*\}_{k=1}^{K}\right)$ of set U.

1: Construct the Voronoi tessellations $\{\mathcal{U}_k\}_{k=1}^{K}$ of U associated with generators $\{z_k\}_{k=1}^{K}$.
2: Compute the centroids $\{z_k^*\}_{k=1}^{K}$ of Voronoi regions $\{\mathcal{U}_k\}_{k=1}^{K}$.
3: For $k = 1, 2, \ldots, K$, if $z_k^* = z_k$, stop; otherwise, let $z_k = z_k^*$ and return to step 1.

3 Pre-classification based SROM

For a given SPDE, we first use the similarity and difference between sample solutions to cluster them, and each class can generate a set of POD basis functions. Then, a pre-classifier is trained by clustering results for assigning unlabelled input, and the corresponding model is reduced by the basis functions of the predicted class. In this section, we propose the t-gCVT clustering method for generating multiple sets of POD basis functions and the naive Bayes pre-classifier based SROM.

3.1 Time-dependent generalized CVT

As mentioned above, the distance in CVT can be extended to other general distances. And from the measurement formula (2.4), we can see that the POD method is to find a subspace that minimizes the expected value of projection distance. Therefore, it is natural to consider combining the POD and CVT methods.

For a given solution u, in order to ensure that the basis functions of a subset after clustering can be used to generate its reduced-order approximation in the entire time domain, the time-dependent distance is defined as

$$\widehat{\mathcal{D}}(u, \Pi^d u) = \|u - \Pi^d u\|_{\mathcal{L}^2([0,T];\, L^2(D))} \tag{3.1}$$

for any d-dimensional subspace Π^d. Given a set of multidimensional subspaces $\{\Pi_k^{d_k}\}_{k=1}^K$, $d_k \in \mathbb{N}^+$, for the solution u of SPDE (2.1), the t-gVT is given as

$$\widehat{\mathcal{U}}_k = \{u \in U_s \mid \widehat{\mathcal{D}}^2(u, \Pi_k^{d_k} u) \le \widehat{\mathcal{D}}^2(u, \Pi_i^{d_i} u) \text{ for all } i \ne k\} \quad k = 1, 2, \ldots, K, \tag{3.2}$$

where U_s denotes the solution space, which is composed of all functions u satisfying system (2.1). Similar to (2.12), the principle of random assignment in the appropriate classes is used to break the deadlock. It is well-known that the traditional CVT method clusters data by trying to separate samples into several classes that have the equal variance in the sense of the given distance. Therefore, the generalized centroid can be naturally defined as the subspace $\widehat{\Pi}_k^{d_k}$ spanned by orthonormal basis functions, which minimizes

$$\widehat{\mathcal{E}}_k^{\text{t-gCVT}}\left(\widehat{\Pi}_k^{d_k}\right) = \mathbb{E}\left[\|u - \widehat{\Pi}_k^{d_k} u\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}\right] \quad k = 1, 2, \ldots, K, \tag{3.3}$$

where $u \in \widehat{\mathcal{U}}_k$ for $k = 1, 2, \ldots, K$. Next, the t-gCVT is derived from the definition of CVT.

Definition 3.1 The t-gVT $(\{\widehat{\mathcal{U}}_k\}_{k=1}^K;\ \{\Pi_k^{d_k}\}_{k=1}^K)$ of the solution space U_s is called t-gCVT if and only if the generator $\Pi_k^{d_k}$ of class $\widehat{\mathcal{U}}_k$ is the corresponding generalized centroid, i.e. $\Pi_k^{d_k} = \widehat{\Pi}_k^{d_k}$, for $k = 1, 2, \ldots, K$.

As can be seen from the above description, in the process of t-gCVT clustering, the calculation of distance (3.1) involves time integral, and the construction of the generalized centroid is difficult because it is required to be optimal over the entire time domain in the sense of expectation. Therefore, the MC method with sample set $\widehat{U} = \{u_i\}_{i=1}^n := \{u(t, \mathbf{x}, \boldsymbol{\xi}_i)\}_{i=1}^n$ is used for the expectation, and the time-sampling snapshots at equal interval are used

to define a modified distance as

$$\widetilde{\mathcal{D}}^2(u_i, \Pi^d u_i) = \sum_{j=1}^{J} \|u(t_j, \mathbf{x}, \boldsymbol{\xi}_i) - \Pi^d u(t_j, \mathbf{x}, \boldsymbol{\xi}_i)\|^2_{L^2(D)} \quad i = 1, 2, \ldots, n, \tag{3.4}$$

where $\{t_j\}_{j=1}^J$ are the corresponding time points of snapshots, $t_0 = 0$ and $t_{j+1} = t_j + \Delta t$ for $j = 0, 1, \ldots, J-1$ with time interval $\Delta t = T/J$. This is equivalent to using the snapshots to approximately calculate the integral with respect to time in distance (3.1), and the scaling factor is Δt. Then the modified t-gVT can be defined as

$$\widetilde{\mathcal{U}}_k = \{u \in \widehat{U} \mid \widetilde{\mathcal{D}}^2(u, \Pi_k^{d_k} u) \leq \widetilde{\mathcal{D}}^2(u, \Pi_i^{d_i} u) \text{ for all } i \neq k\} \quad k = 1, 2, \ldots, K, \tag{3.5}$$

and the modified generalized centroid $\widetilde{\Pi}_k^{d_k} = \text{span}\{\phi_1^k, \ldots, \phi_{d_k}^k\}$ minimizes

$$\widetilde{\mathcal{E}}_k^{\text{t-gCVT}}\left(\widetilde{\Pi}_k^{d_k}\right) = \sum_{u \in \widetilde{\mathcal{U}}_k} \sum_{j=1}^{J} \|u(t_j, \mathbf{x}, \boldsymbol{\xi}) - \widetilde{\Pi}_k^{d_k} u(t_j, \mathbf{x}, \boldsymbol{\xi})\|^2_{L^2(D)} \quad k = 1, 2, \ldots, K. \tag{3.6}$$

Denote the cardinality of $\widetilde{\mathcal{U}}_k$ as n_k, which satisfies $\sum_{k=1}^K n_k = n$. According to the optimality of POD, for $k = 1, 2, \ldots, K$, the modified generalized centroid $\widetilde{\Pi}_k^{d_k}$ is actual the subspace spanned by the POD basis functions, which are generated by the snapshots of set $\widetilde{\mathcal{U}}_k$.

If the approximate error of the time integral is negligible, that is,

$$\widehat{D}^2(u, \Pi^d u) = \Delta t \widetilde{D}^2(u, \Pi^d u) \tag{3.7}$$

holds for any given subspace Π^d. Then the following inequality is known from the relationship between the minimum value of the expected value and the expectation of the minimum value

$$\min \mathbb{E}\left[\widehat{D}^2(u, \Pi^d u)\right] \geq \Delta t \mathbb{E}\left[\min \widetilde{D}^2(u, \Pi^d u)\right]. \tag{3.8}$$

Therefore, $\{\widehat{\mathcal{E}}_k^{\text{t-gCVT}}\}_{k=1}^K$ and $\{\widetilde{\mathcal{E}}_k^{\text{t-gCVT}}\}_{k=1}^K$ satisfy

$$\min \widehat{\mathcal{E}}_k^{\text{t-gCVT}} \geq \frac{\Delta t}{n_k} \min \widetilde{\mathcal{E}}_k^{\text{t-gCVT}} \qquad k = 1, 2, \ldots, K \tag{3.9}$$

by using the MC method with n_k samples of set $\widetilde{\mathcal{U}}_k$ to estimate the right-hand side of inequality (3.8).

Similar to Definition 3.1, the definition of modified t-gCVT is given as follows.

Definition 3.2 The modified t-gVT $(\{\widetilde{\mathcal{U}}_k\}_{k=1}^K;\ \{\Pi_k^{d_k}\}_{k=1}^K)$ of the set $\widehat{U}$ is called modified t-gCVT if and only if the generator $\Pi_k^{d_k}$ of set $\widetilde{\mathcal{U}}_k$ is the corresponding generalized centroid, i.e. $\Pi_k^{d_k} = \widetilde{\Pi}_k^{d_k}$, for $k = 1, 2, \ldots, K$. And the POD basis functions $\{\phi_j^k\}_{j=1}^{d_k}$ corresponding to the generalized centroid $\widetilde{\Pi}_k^{d_k}$ of modified t-gCVT are called its subclass bases or cluster-based POD (CPOD) basis functions.

It can be seen from the above definition that the modified t-gCVT of set $\widehat{U}$ minimizes the error

$$\widetilde{\mathcal{E}}^{\text{t-gCVT}} = \sum_{k=1}^{K} \widetilde{\mathcal{E}}_k^{\text{t-gCVT}} \left(\widetilde{\Pi}_k^{d_k}\right), \tag{3.10}$$

and the minimum value is

$$\widetilde{\mathcal{E}}^{\text{t-gCVT}} = \sum_{k=1}^{K} Jn_k \sum_{j=d_k+1}^{Jn_k} \sigma_j^k, \tag{3.11}$$

where $\{\sigma_j^k\}_{j=1}^{Jn_k}$ are the eigenvalues of correlation matrix R associated with set $\widetilde{\mathcal{U}}_k$, as difined in (2.10). Here, (3.11) is referred to as "modified t-gCVT energy", and

$$\nu_k = \sum_{j=1}^{d_k} \sigma_j^k \Big/ \sum_{j=1}^{Jn_k} \sigma_j^k \qquad k = 1, 2, \ldots, K \tag{3.12}$$

is called the energy ratio of CPOD bases $\{\phi_j^k\}_{j=1}^{d_k}$.

To reduce the complexity of model construction, the modified t-gCVT is used in the following processes, and its structure is shown in Figure 1. Note that the modified t-gCVT is reduced to the standard snapshot-based POD method when $K = 1$, and the number of CPOD basis functions $\{\phi_j^k\}_{j=1}^{d_k}$ is not neccessarily equal for $k = 1, 2, \ldots, K$.

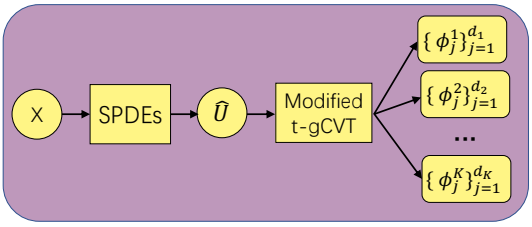

Fig. 1: The framework of modified t-gCVT method

Remark 3.3 When the modified t-gCVT $(\{\widetilde{\mathcal{U}}_k\}_{k=1}^K;\ \{\widetilde{\Pi}_k^{d_k}\}_{k=1}^K)$ of set $\widehat{U}$ is known, we can naturally cluster the inputs $\{\boldsymbol{\xi}_i\}_{i=1}^n$ according to the clustering results of data

$\widehat{U}$. Namely, the image space Γ of input $\boldsymbol{\xi}$ can be divided into $\{\Gamma_k\}_{k=1}^K$, which satisfies $\Gamma_i \cap \Gamma_j = \emptyset$ if $i \neq j$, $\Gamma_k \subset \Gamma$ for $k = 1, 2, \ldots, K$ and $\bigcup_{k=1}^K \Gamma_k = \Gamma$. If $u(t, \mathbf{x}, \boldsymbol{\xi}) \in \widetilde{\mathcal{U}}_k$, then the corresponding input, $\boldsymbol{\xi} \in \Gamma$, is belonging to Γ_k, i.e.,

$$\Gamma_k = \{\boldsymbol{\xi} \in \Gamma \mid u(t, \mathbf{x}, \boldsymbol{\xi}) \in \widetilde{\mathcal{U}}_k\} \qquad k = 1, 2, \ldots, K, \tag{3.13}$$

where k is called the class label of $\boldsymbol{\xi}$.

The details of using the modified t-gCVT method to generate the CPOD basis functions are given in Algorithm 3.

Algorithm 3 The modified t-gCVT clustering method for generating CPOD basis functions

Input: set $\widehat{U} = \{u(t, \mathbf{x}, \boldsymbol{\xi}_i)\}_{i=1}^n$, a positive integer K, dimensions $\{d_k\}_{k=1}^K$, step size Δt.

Output: modified t-gCVT $(\{\widetilde{\mathcal{U}}_k\}_{k=1}^K;\ \{\widetilde{\Pi}_k^{d_k}\}_{k=1}^K)$ of $\widehat{U}$, and K groups of CPOD basis functions $\{\{\phi_j^1\}_{j=1}^{d_1}, \ldots, \{\phi_j^K\}_{j=1}^{d_K}\}$.

1: Select a set of initial generalized generators $\{\Pi_k^{d_k}\}_{k=1}^K$ with dimensions $\{d_k\}_{k=1}^K$.
2: Construct the modified t-gVT $\{\widetilde{\mathcal{U}}_k\}_{k=1}^K$ of $\widehat{U}$ associated with $\{\Pi_k^{d_k}\}_{k=1}^K$.
3: From $\{\widetilde{\mathcal{U}}_k\}_{k=1}^K$ and step size Δt, determine the snapshot sets $\{\mathcal{W}_k\}_{k=1}^K$ defined in (2.8).
4: Generate K groups CPOD basis functions $\{\{\phi_j^1\}_{j=1}^{d_1}, \ldots, \{\phi_j^K\}_{j=1}^{d_K}\}$ defined in (2.9) by solving eigenvalue problems (2.10) associated with $\{\mathcal{W}_k\}_{k=1}^K$
5: For $k = 1, 2, \ldots, K$, let $\widetilde{\Pi}_k^{d_k} = \text{span}(\phi_1^k, \ldots, \phi_{d_k}^k)$, if $\Pi_k^{d_k} = \widetilde{\Pi}_k^{d_k}$, stop; otherwise, let $\Pi_k^{d_k} = \widetilde{\Pi}_k^{d_k}$ and return to step 2.

3.2 Naive Bayes pre-classifier and pre-classification based SROM

Since the modified t-gCVT method is to cluster the spatio-temporal function u, then for a given u, a set of suitable CPOD basis functions can be used to calculate its reduced-order approximation in the whole time interval. In modified t-gCVT, the set $\widetilde{\mathcal{U}}_k$ with the highest similarity to the function u is called its best-matched set, and the corresponding CPOD bases are called the best-matched bases. In general, the reduced-order approximation generated by the best-matched bases is better than the standard POD approximation with the same degree of freedom (DoF). This is because that the samples in the same class are similar after clustering, then the same number of basis functions can capture more useful information, which is beneficial for the reconstruction of function u. That is to say, if we know the best-matched bases of a given function, the accuracy of its reduced-order approximation can be improved compared with the standard POD method. Note that the spatio-temporal

function $u(t, \mathbf{x}, \boldsymbol{\xi})$ is determined by the random input $\boldsymbol{\xi}$, and our aim is to construct a SROM such that the approximate solution can be obtained for any given input $\boldsymbol{\xi}$. Therefore, a pre-classifier is constructed here to select the best-matched bases from the perspective of random input.

In this paper, the naive Bayes pre-classifier based on Bayes' theorem and the assumption of feature condition independence is adopted. For a given integer $K \geq 1$, the image space Γ is divided into disjoint subspace set $\{\Gamma_k\}_{k=1}^K$ as introduced in Remark 3.3. Suppose $\boldsymbol{\gamma}$ is a random vector defined on the input space $\Gamma \subset \mathbb{R}^p$ composed of p-dimensional vectors. Its realization, also known as the feature vector, is denoted as $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_p]^\top \in \Gamma$. Let ι be a random variable defined on the class label set $\mathcal{L} = \{1, \ldots, K\}$. Its realization, also known as class label, is denoted as $k \in \mathcal{L}$. Let $X = \{\boldsymbol{\xi}_i\}_{i=1}^n$ be the independent and identically distributed (i.i.d.) input set of the given data $\widehat{U}$, and $\{\iota_i\}_{i=1}^n$ be the corresponding class labels obtained by the modified t-gCVT method, then the training data set is given as

$$\mathbb{D} = \{(\boldsymbol{\xi}_1, \iota_1), \ldots, (\boldsymbol{\xi}_n, \iota_n)\}. \tag{3.14}$$

Denote the prior probability distributions

$$\mathbb{P}(\iota = k) = \pi_k \qquad k = 1, 2, \ldots, K, \tag{3.15}$$

and conditional probability distributions

$$\mathbb{P}(\boldsymbol{\gamma} = \boldsymbol{\xi} | \iota = k) = \prod_{i=1}^{p} \mathbb{P}(\gamma_i = \xi_i | \iota = k) = f_k(\boldsymbol{\xi}) \qquad k = 1, 2, \ldots, K \tag{3.16}$$

as

$$\pi_k = \frac{n_k}{n} \qquad k = 1, 2, \ldots, K \tag{3.17}$$

and

$$f_k(\boldsymbol{\xi}) = \prod_{i=1}^{p} f_k(\xi_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i^{(k)}} \exp\left(-\frac{|\xi_i - \mu_i^{(k)}|^2}{2(\sigma_i^{(k)})^2}\right) \quad k = 1, 2, \ldots, K, \tag{3.18}$$

respectively. Here, the means $\{\mu_i^{(k)}\}$ and variances $\{\sigma_i^{(k)}\}$ can be estimated by

$$\mu_i^{(k)} = \frac{1}{n_k} \sum_{\boldsymbol{\xi} \in \Gamma_k} \xi_i \qquad i = 1, 2, \ldots, p, \ k = 1, 2, \ldots, K, \tag{3.19}$$

$$\sigma_i^{(k)} = \frac{1}{n_k - 1} \sum_{\boldsymbol{\xi} \in \Gamma_k} \left(\xi_i - \mu_i^{(k)}\right)^2 \qquad i = 1, 2, \ldots, p, \ k = 1, 2, \ldots, K. \tag{3.20}$$

According to the Bayes' theorem, the posterior probability has form

$$\mathbb{P}(\iota = k|\boldsymbol{\gamma} = \boldsymbol{\xi}) = \frac{\pi_k f_k(\boldsymbol{\xi})}{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{\xi})}. \tag{3.21}$$

The principle of naive Bayes pre-classifier is to assign input to the class with the largest posterior probability, that is, input $\boldsymbol{\xi}$ is assigned to the subspace Γ_k if

$$k = \arg\max_{1 \le i \le K} \mathbb{P}(\iota = i|\boldsymbol{\gamma} = \boldsymbol{\xi}). \tag{3.22}$$

The denominator of (3.21) is a fixed constant for a given $\boldsymbol{\xi}$, so (3.22) is equivalent to

$$k = \arg\max_{1 \le i \le K} \pi_i f_i(\boldsymbol{\xi}). \tag{3.23}$$

If the result in (3.23) is not unique, a random assignment is used to break the tie. Here, k is the predicted label of input $\boldsymbol{\xi}$, and the corresponding $\widetilde{\mathcal{U}}_k$ and $\{\phi_j^k\}_{j=1}^{d_k}$ are called the predicted best-matched set and predicted best-matched bases of solution $u(t, \mathbf{x}, \boldsymbol{\xi})$, respectively.

The naive Bayes classifier is based on the assumption of normality and independence of variables, which will affect the accuracy of the algorithm to a certain extent. But this algorithm is easy to implement and has high learning and prediction efficiency. Therefore, it is still one of the popular classification tools.

When the naive Bayes pre-classifier assigns an unlabelled input $\boldsymbol{\xi}$ to the subspace Γ_k, that is to say, the probability of $\boldsymbol{\xi} \in \Gamma_k$ is the largest, then the continuity of the input-output mapping shows that its solution u is most likely to belong to the set $\widetilde{\mathcal{U}}_k$. Therefore, it is feasible to use k-th group CPOD bases $\{\phi_j^k\}_{j=1}^{d_k}$ of modified t-gCVT to evaluate the corresponding model, and the approximation of solution u is given by

$$\widetilde{u}^K(t, \mathbf{x}, \boldsymbol{\xi}) = \sum_{j=1}^{d_k} \alpha_j(t, \boldsymbol{\xi}) \phi_j^k(\mathbf{x}), \tag{3.24}$$

where $\{\alpha_j\}_{j=1}^{d_k}$ can be obtained by solving the following reduced system

$$\left\langle F\left(\sum_{j=1}^{d_k} \alpha_j(t, \boldsymbol{\xi}) \phi_j^k(\mathbf{x});\ \boldsymbol{\xi}\right), \phi_i^k(\mathbf{x}) \right\rangle = 0 \qquad i = 1, 2, \ldots, d_k. \tag{3.25}$$

We call the method of combining CPOD basis functions and naive Bayes pre-classifier to construct SROM as the *CPOD-NB method*, and $\widetilde{u}^K$ defined in (3.24) is the CPOD-NB reduced-order approximation of solution u with the number of clusters K. The structure of the model reduction based on CPOD-NB method is shown in Figure 2.

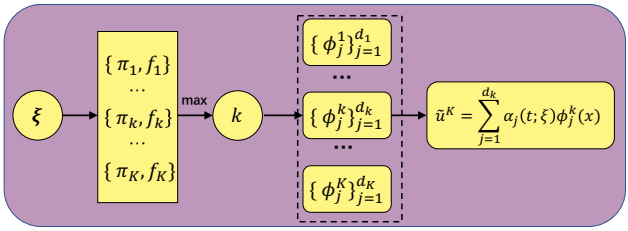

Fig. 2: The framework of model reduction based on CPOD-NB method

So far, the modified t-gCVT method and pre-classification based SROM have been introduced, and the details of CPOD-NB method for model reduction are described in Algorithm 4.

Algorithm 4 SROM based on CPOD-NB method

Input: input set $X = \{\boldsymbol{\xi}_i\}_{i=1}^n$, a positive integer K, dimensions $\{d_k\}_{k=1}^K$, step size Δt.

Output: CPOD-NB approximate solution $\widetilde{u}^K$ of new input $\boldsymbol{\xi}$.

1: Generate data set $\widehat{U}$ by solving system (2.1) with inputs X.
2: Obtain the modified t-gCVT $(\{\widetilde{\mathcal{U}}_k\}_{k=1}^K;\ \{\widetilde{\Pi}_k^{d_k}\}_{k=1}^K)$ of $\widehat{U}$ and K groups of CPOD basis functions $\{\{\phi_j^1\}_{j=1}^{d_1}, \ldots, \{\phi_j^K\}_{j=1}^{d_K}\}$ by using Algirithm 3.
3: For $i = 1, 2, \ldots, n$, if $u(t, \mathbf{x}, \boldsymbol{\xi}_i) \in \widetilde{\mathcal{U}}_k$, then denote the label of $\boldsymbol{\xi}_i$ as $\iota_i = k$, where $k \in \{1, 2, \ldots, K\}$.
4: Use the input set X and the labels $\{\iota_i\}_{i=1}^n$ to form the training data set $\mathbb{D}$, then learn the prior probability distributions $\{\pi_k\}_{k=1}^K$ and conditional probability density functions $\{f_k\}_{k=1}^K$.
5: For a given new input $\boldsymbol{\xi}$, compute the values of $\{\pi_k, f_k(\boldsymbol{\xi})\}_{k=1}^K$, then assign $\boldsymbol{\xi}$ to Γ_k if (3.23) holds.
6: Obtain the reduced states $\{\alpha_j\}_{j=1}^{d_k}$ by solving the system (3.25) with k-th group CPOD basis functions $\{\phi_j^k\}_{j=1}^{d_k}$, then the CPOD-NB approximate solution $\widetilde{u}^K$ of $\boldsymbol{\xi}$ can be represented as (3.24).

Remark 3.4 For a given input $\boldsymbol{\xi}$, in the CPOD-NB method, we hope to find the set of CPOD basis functions such that the error between its finite element solution and the reduced-order solution is the smallest. Therefore, the true label of input $\boldsymbol{\xi}$ can be defined as

$$i = \arg\min_{1\le k\le K} \left\| u(t, \mathbf{x}, \boldsymbol{\xi}) - \widetilde{\Pi}_k^{d_k} u(t, \mathbf{x}, \boldsymbol{\xi}) \right\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}, \tag{3.26}$$

and the corresponding $\widetilde{\mathcal{U}}_i$ and $\{\phi_j^i\}_{j=1}^{d_i}$ are called the true best-matched set and true best-matched bases of solution $u(t, \mathbf{x}, \boldsymbol{\xi})$, respectively.

4 Error estimation

In this section, we first give the error estimation of the SROM based on CPOD-NB method, and then introduce the estimation method of error rate of the naive Bayes pre-classifier.

4.1 Error estimation of CPOD-NB based SROM

In order to characterize the validity of the CPOD-NB model, the error between the full finite element solution u and the CPOD-NB approximate solution $\widetilde{u}^K$ is defined as

$$\widetilde{\mathcal{E}}_K = \mathbb{E}\left[\|u - \widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}\right] \tag{4.1}$$

and

$$\widetilde{\mathcal{V}}_K = \mathbb{V}\left[\|u - \widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}\right], \tag{4.2}$$

where $\mathbb{V}[\cdot]$ represents the variance.

The error estimation of the CPOD-NB reduced-order solution is given in following theorem.

Theorem 4.1 *In the naive Bayes pre-classifier, if the random input $\boldsymbol{\xi}$ can always get the true label with the maximum posterior probability, then there exist constants $C_1, C_2 > 0$, such that with probability close to one, the space-time $L^2(D)$-norm error $\widetilde{\mathcal{E}}_K$ between the finite element solution u and the CPOD-NB approximate solution $\widetilde{u}^K$ satisfies*

$$\widetilde{\mathcal{E}}_K \le \sum_{k=1}^{K}\left(\frac{Tn_k}{n}\sum_{j=d_k+1}^{Jn_k}\sigma_j^k\right) + C_1\sqrt{\widetilde{\mathcal{V}}_K/n} + C_2\frac{T\Delta t}{2}, \tag{4.3}$$

where C_2 depends on the regularity of $\|u(t) - \widetilde{u}^K(t)\|^2_{L^2(D)}$, while constant C_1 is universal.

Proof By using the MC method, the error can be estimated by

$$\widetilde{\mathcal{E}}_K = \frac{1}{n}\sum_{i=1}^{n}\|u(t,\mathbf{x},\boldsymbol{\xi}_i) - \widetilde{u}^K(t,\mathbf{x},\boldsymbol{\xi}_i)\|^2_{\mathcal{L}^2([0,T];\, L^2(D))} + \widetilde{\mathcal{E}}_s,$$

where $\widetilde{\mathcal{E}}_s$ denotes statistical error and satisfies

$$\widetilde{\mathcal{E}}_s \sim N(0, \widetilde{\mathcal{V}}_K/n)$$

according to the central limit theorem. For a constant $C_1 \ge 1.65$, the inequality

$$|\widetilde{\mathcal{E}}_s| \le C_1\sqrt{\widetilde{\mathcal{V}}_K/n}$$

can hold with probability close to 1. Then using data $\widehat{U}$ and its clustering results, the following can be obtained

$$\widetilde{\mathcal{E}}_K - \widetilde{\mathcal{E}}_s = \frac{1}{n}\sum_{k=1}^{K}\sum_{u\in\widetilde{\mathcal{U}}_k}\|u - \widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}.$$

The time-sampling snapshots at equal interval are used to approximate the time integral, that is

$$\|u-\widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\,L^2(D))}=\Delta t\sum_{j=1}^{J}\|u(t_j)-\widetilde{u}^K(t_j)\|^2_{L^2(D)}+R[u,\widetilde{u}^K],$$

where time step $\Delta t = T/J$, $t_0 = 0$ and $t_{j+1} = t_j + \Delta t$ for $j = 0, 1, \ldots, J-1$. $R[u, \widetilde{u}^K]$ is the residual of the approximation which depends on the regularity of $f(t;\, u) := \|u(t) - \widetilde{u}^K(t)\|^2_{L^2(D)}$ and satisfies

$$R[u,\widetilde{u}^K]=\frac{T\Delta t}{2}f'(\eta;\,u)$$

for some $\eta \in (0, T)$. Therefore,

$$\widetilde{\mathcal{E}}_K-\widetilde{\mathcal{E}}_s=\frac{\Delta t}{n}\sum_{k=1}^{K}\sum_{u\in\widetilde{\mathcal{U}}_k}\sum_{j=1}^{J}\|u(t_j)-\widetilde{u}^K(t_j)\|^2_{L^2(D)}+\frac{1}{n}\sum_{k=1}^{K}\sum_{u\in\widetilde{\mathcal{U}}_k}\frac{T\Delta t}{2}f'(\eta;\,u).$$

Let

$$C_2=\max_{u\in\widehat{U},\eta\in(0,T)}|f'(\eta;\,u)|,$$

then according to the energy (3.11)

$$\widetilde{\mathcal{E}}_K-\widetilde{\mathcal{E}}_s\le\sum_{k=1}^{K}\left(\frac{Jn_k\Delta t}{n}\sum_{j=d_k+1}^{Jn_k}\sigma_j^k\right)+C_2\frac{T\Delta t}{2}$$

holds, which completes the proof. □

4.2 Error rate estimation of the naive Bayes pre-classifier

In general, classification rules have their error rate. When the Bayes classifier with the maximum posterior decision rule is used to classify the problem with known conditional probability density functions and prior probability distributions, its error rate should be fixed. Next, we consider the error rate estimation of the naive Bayesian pre-classifier.

According to the statistical decision theory [35], denote the discriminant functions as

$$g_k(\boldsymbol{\xi})=\pi_k f_k(\boldsymbol{\xi})\qquad k=1,2,\ldots,K, \tag{4.4}$$

and their decision regions are defined by

$$\widetilde{\Gamma}_k=\{\boldsymbol{\xi}\in\Gamma\mid g_k(\boldsymbol{\xi})>g_i(\boldsymbol{\xi})\text{ for }i=1,2,\ldots,K,\ i\neq k\}\qquad k=1,2,\ldots,K. \tag{4.5}$$

Then the decision surface between regions $\widetilde{\Gamma}_i$ and $\widetilde{\Gamma}_j$ is given as

$$\mathcal{S}_{ij}=\{\boldsymbol{\xi}\in\Gamma\mid g_i(\boldsymbol{\xi})=g_j(\boldsymbol{\xi}),\ i\neq j\}\qquad i,j=1,2,\ldots,K. \tag{4.6}$$

Note that the decision region set $\{\widetilde{\Gamma}_k\}_{k=1}^K$ is also a partition of the feature space Γ. Although we hope that it is consistent with the segmentation $\{\Gamma_k\}_{k=1}^K$ in the modified t-gCVT so that the input samples can always be assigned to the best subspace with the maximum posterior probability, it is difficult to

achieve in practice due to the defects of the classifier itself and the lack of data. Therefore, it is necessary to study the error rate of classifier.

According to the classification rules of naive Bayes, its error rate $\mathbb{P}(e)$ is the probability of assigning sample that belongs to subspace Γ_k to other subspace Γ_i, where $i, k = 1, 2, \ldots, K$ and $i \neq k$. That is

$$\mathbb{P}(e) = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbb{P}(\boldsymbol{\xi} \in \widetilde{\Gamma}_i, \iota = k) = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} \mathbb{P}(\boldsymbol{\xi} \in \widetilde{\Gamma}_i | \iota = k) \mathbb{P}(\iota = k) = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} \pi_k \mathbb{P}_{ki}(e), \tag{4.7}$$

where

$$\mathbb{P}_{ki}(e) = \mathbb{P}(\boldsymbol{\xi} \in \widetilde{\Gamma}_i | \iota = k) = \int_{\widetilde{\Gamma}_i} \mathbb{P}(\boldsymbol{\gamma} = \boldsymbol{\xi} | \iota = k) d\boldsymbol{\xi} = \int_{\widetilde{\Gamma}_i} f_k(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{4.8}$$

Then the correct rate of the classifier takes the form

$$\mathbb{P}(c) = 1 - \mathbb{P}(e) = \sum_{k=1}^{K} \pi_k \mathbb{P}_{kk}(e). \tag{4.9}$$

For high-dimensional stochastic problems, it is difficult to determine the decision regions $\{\widetilde{\Gamma}_k\}$ and the decision surfaces $\{S_{ij}\}$, so the calculation of integrals (4.8) is a huge challenge. Here, a more practical method can be used to estimate the error rate for testing the performance of the classifier.

A test set $\mathbb{T} = \{\boldsymbol{\xi}_i\}_{i=1}^{N}$ with size N is randomly selected from the feature space Γ, and its components are mutually independent and independent of the training data $\mathbb{D}$. Let the total number of samples in Γ_k be N_k for $k = 1, 2, \ldots, K$, which satisfy $\sum_{k=1}^{K} N_k = N$. The number of samples belonging to subspace Γ_k that are misjudged into subspace Γ_i is denoted as n_{ki} for $k, i = 1, 2, \ldots, K$ and $i \neq k$. Obviously, n_{ki} is a discrete random variable that obeys a binomial distribution and satisfies

$$\mathbb{P}(n_{ki}) = C_{N_k}^{n_{ki}} \left[\mathbb{P}_{ki}(e)\right]^{n_{ki}} \left[1 - \mathbb{P}_{ki}(e)\right]^{N_k - n_{ki}}, \tag{4.10}$$

where $C_{N_k}^{n_{ki}} = \frac{N_k!}{n_{ki}!(N_k - n_{ki})!}$. By solving

$$\frac{\partial \ln \mathbb{P}(n_{ki})}{\partial \mathbb{P}_{ki}(e)} = 0 \tag{4.11}$$

can obtain the maximum likelihood estimation of $\mathbb{P}_{ki}(e)$ as

$$\widehat{\mathbb{P}}_{ki}(e) = \frac{n_{ki}}{N_k}, \tag{4.12}$$

which is also a random variable, and the mean has form

$$\mathbb{E}\left[\widehat{\mathbb{P}}_{ki}(e)\right] = \frac{\mathbb{E}\left[n_{ki}\right]}{N_k} = \mathbb{P}_{ki}(e). \tag{4.13}$$

Therefore, $\widehat{\mathbb{P}}_{ki}(e)$ is an unbiased estimate of $\mathbb{P}_{ki}(e)$, and further an unbiased estimate of $\mathbb{P}(e)$ can be obtained as

$$\widehat{\mathbb{P}}(e) = \sum_{k=1}^{K} \sum_{\substack{i=1 \\ i \neq k}}^{K} \pi_k \widehat{\mathbb{P}}_{ki}(e). \tag{4.14}$$

In numerical experiments of this work, we use formula (4.14) to estimate the error rate of the naive Bayes pre-classifier.

5 Stochastic Navier-Stokes equations

In this work, we use the proposed CPOD-NB based SROM to deal with stochastic flow over a backward-facing step[36] described as

$$\mathbf{u}_t - \frac{1}{Re}\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla P = 0 \qquad (0,T]\times D, \tag{5.1}$$

$$\nabla\cdot\mathbf{u} = 0 \qquad (0,T]\times D, \tag{5.2}$$

where Re is the Reynolds number of the fluid, $\mathbf{u}(t,\mathbf{x}) = (u_1,u_2)^\top$ and $P(t,\mathbf{x})$ denote the velocity and pressure fields, respectively. The boundary of physical domain D is denoted by ∂D, which consists of six parts as depicted in Figure 3. For $t\in(0,T]$, the boundary conditions are given by

$$\mathbf{u} = (u_{\text{in}},0)^\top \qquad \text{on } \partial D_i, \tag{5.3}$$

$$\mathbf{u} = (0,0)^\top \qquad \text{on } \partial D_t \cup \partial D_b \cup \partial D_d \cup \partial D_c, \tag{5.4}$$

$$P\mathbf{n} - \frac{1}{Re}\frac{\partial\mathbf{u}}{\partial\mathbf{n}} = (0,0)^\top \qquad \text{on } \partial D_o, \tag{5.5}$$

and the initial velocity field satisfies

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(x,y) = \begin{cases} u_{\text{in}}(0,\mathbf{x}) & \text{on } \partial D_i, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

Assume that the fluid can be injected along ∂D_i, so $u_{\text{in}}\geq 0$ is required. Further assume that the injected fluid contains uncertainties. Thus, for a properly defined probability space $(\Omega,\mathcal{F},\mathbb{P})$, u_{in} can be modelled with random variable $\omega\in\Omega$ as

$$u_{\text{in}} = A(t,\boldsymbol{\xi}(\omega))h(y), \tag{5.7}$$

where A is a time-dependent parameter that determines the strength of the parabolic inflow velocity profile $h(y)$. For the sake of simplicity, denote $A(t, \boldsymbol{\xi}(\omega))$ as $A(t, \boldsymbol{\xi})$ or $A(t;\, \omega)$.

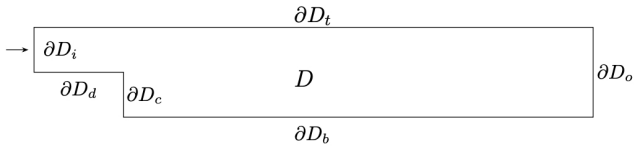

Fig. 3: Physical domain D of Navier-Stokes equation

5.1 Full discrete and Newton linearization

In this paper, finite element method, θ-scheme and Newton's method are used for spatial discretization, time discretization, and the linearization of nonlinear convective term, respectively.

Let $\mathcal{T}^h$ be a shape-regular triangular finite element mesh of domain D, which is parameterized by mesh width $h = \max_{G\in\mathcal{T}^h} \mathrm{diam}(G)$, where G is a typical finite element in the triangulation $\mathcal{T}^h$. The finite element mesh used in this work is shown in Figure 4. For vector valued function $\mathbf{u}$, define the following finite element spaces

$$\begin{aligned}
&\mathbf{V}^h = \{\mathbf{v}^h = (v_1^h, v_2^h)^\top : v_i^h \in C^0(\bar{D}), v_i^h|_G \in \mathcal{P}^2 \text{ for any } G \in \mathcal{T}^h, i = 1, 2\},\\
&\mathbf{V}_0^h = \{\mathbf{v}^h \in \mathbf{V}^h : v_i^h = 0 \text{ on } \partial D \setminus \partial D_o \text{ for } i = 1, 2\},\\
&Q^h = \{q^h : q^h \in C^0(\bar{D}), q^h|_G \in \mathcal{P}^1 \text{ for any } G \in \mathcal{T}^h\},
\end{aligned}$$

where $\mathcal{P}^r$ denotes the polynomial space with degree less than or equal to r, $r \in \mathbb{N}^+$. The Taylor-Hood finite element spaces are considered in our computation, i.e. quadratic finite element space for velocity field $\mathbf{u}$ and linear finite element space for pressure field P.

Let $\tau_m = \{t_i\}_{i=0}^m$ be a partition of $[0, T]$ with equal interval $Dt = T/m$, where $t_0 = 0$ and $t_i = t_{i-1} + Dt$ for $i = 1, 2, \ldots, m$. Then for $i = 0, 1, \ldots, m-1$, the linearized full discrete weak formulation of system (5.1)-(5.5) is given as: find $\mathbf{u}_\theta^{i+1} \in \mathbf{V}^h$ and $P_\theta^{i+1} \in Q^h$ such that

$$\begin{aligned}
\frac{1}{\theta\Delta t}\int_D \left(\mathbf{u}_\theta^{i+1} - \mathbf{u}^i\right)\mathbf{v}d\mathbf{x} + \frac{1}{Re}\int_D \nabla\mathbf{u}_\theta^{i+1} : \nabla\mathbf{v}d\mathbf{x} + \int_D \mathbf{u}_\theta^{i+1}\cdot\nabla\mathbf{u}^i\cdot\mathbf{v}d\mathbf{x}\\
+ \int_D \mathbf{u}^i\cdot\nabla\mathbf{u}_\theta^{i+1}\cdot\mathbf{v}d\mathbf{x} - \int_D P_\theta^{i+1}\nabla\cdot\mathbf{v}d\mathbf{x} = \int_D \mathbf{u}^i\cdot\nabla\mathbf{u}^i\cdot\mathbf{v}d\mathbf{x}
\end{aligned} \tag{5.8}$$

$$\int_D q\nabla\cdot\mathbf{u}_\theta^{i+1}d\mathbf{x} = \begin{cases} 0 & \text{if } i > 0,\\ (1-\theta)\int_D q\nabla\cdot\mathbf{u}^0 d\mathbf{x} & \text{if } i = 0,\end{cases}$$

for any test functions $\mathbf{v}^h \in \mathbf{V}_0^h$ and $q^h \in Q^h$. Here, $\mathbf{u}^i = \mathbf{u}(t_i, \mathbf{x})$ and θ is taken as $\frac{1}{2}$. By solving linear system (5.8), the pair $(\mathbf{u}^{i+1}, P^{i+1})$ can be recovered from

$$\mathbf{u}^{i+1} = 2\mathbf{u}_\theta^{i+1} - \mathbf{u}^i \qquad \text{and} \qquad P^{i+1} = 2P_\theta^{i+1} - P^i. \tag{5.9}$$

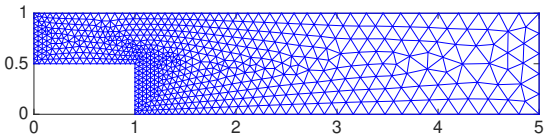

Fig. 4: Finite element mesh with $h = 0.2$, 1279 triangles and 703 vertices

5.2 Modified velocity field

Here, instead of the original finite element solution $\mathbf{u}$, a CPOD-NB model is constructed for the modified velocity field with homogeneous Dirichlet boundaries.

Denote the solutions of the steady-state version of Navier-Stokes system (5.1)-(5.5) with constant strengths $A = a_1$ and $A = a_2$ in inflow velocity u_{in} as $\mathbf{u}_{a_1}$ and $\mathbf{u}_{a_2}$, respectively. Let

$$\mathbf{w} = \frac{\mathbf{u}_{a_1} - \mathbf{u}_{a_2}}{a_1 - a_2}, \tag{5.10}$$

and denote the average of the velocity field as

$$\overline{\mathbf{u}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{J}\sum_{j=1}^{J}\left(\mathbf{u}(t_j, \mathbf{x}, \boldsymbol{\xi}_i) - A(t_j, \boldsymbol{\xi}_i)\mathbf{w}(\mathbf{x})\right)\right). \tag{5.11}$$

Then the modified state is given by

$$\mathbf{v}(t, \mathbf{x}, \boldsymbol{\xi}) = \mathbf{u}(t, \mathbf{x}, \boldsymbol{\xi}) - \overline{\mathbf{u}}(\mathbf{x}) - A(t, \boldsymbol{\xi})\mathbf{w}(\mathbf{x}), \tag{5.12}$$

which satisfies $\mathbf{v} = 0$ on $\partial D \setminus \partial D_o$.

Using the modified t-gCVT method for modified state $\mathbf{v}$, we can obtain K sets of basis functions $\{\{\phi_j^1(\mathbf{x})\}_{j=1}^{d_1}, \ldots, \{\phi_j^K(\mathbf{x})\}_{j=1}^{d_K}\}$. If the class label of a given input $\boldsymbol{\xi}$ is k, the original system (5.8) can be reduced to a d_k-dimensional ordinary differential equations by using $\{\phi_j^k(\mathbf{x})\}_{j=1}^{d_k}$, then the reduced states $\{\alpha_j(t, \boldsymbol{\xi})\}_{j=1}^{d_k}$ can be calculated by Runge-Kutta method, finally the approximation of the original velocity field can be represented as

$$\mathbf{u}(t, \mathbf{x}, \boldsymbol{\xi}) = \overline{\mathbf{u}}(\mathbf{x}) + A(t, \boldsymbol{\xi})\mathbf{w}(\mathbf{x}) + \sum_{j=1}^{d_k}\alpha_j(t, \boldsymbol{\xi})\phi_j^k(\mathbf{x}). \tag{5.13}$$

6 Numerical experiments

To illustrate the feasibility and effectiveness of the proposed CPOD-NB model, we provide comparisons with the standard POD method (i.e. $K = 1$). All computations were performed using MATLAB R2017a on a personal computer with 2.3 GHz CPU and 256 GB RAM.

In our computation, the physical domain D and its triangulation used in the finite element method are shown in the Figure 4. The Reynolds number Re is taken as 500. The finite element solutions of steady-state version of Navier-Stokes system associated with $a_1 = 2$ and $a_2 = 1$ are used to generate the modified state, as defined in (5.12). The time interval $[0, T]$, $T = 2$, is divided by the time step $Dt = 1/200$, and the modified snapshots are obtained at each time point for computing the modified distance, i.e. $\Delta t = Dt$. The parabolic profile $h(y)$ of inflow velocity has form

$$h(y) = (1 - y)(y - 0.5). \tag{6.1}$$

Let the random input of system (5.1)-(5.6) be the time-discrete form of strength $A(t;\, \omega)$, i.e.

$$\boldsymbol{\xi}(\omega) = [\xi_1(\omega), \ldots, \xi_{m+1}(\omega)]^\top = [A(t_0;\, \omega), A(t_1;\, \omega), \ldots, A(t_m;\, \omega)]^\top, \tag{6.2}$$

where $t_0 = 0$, $t_j = t_{j-1} + Dt$ for $j = 1, 2, \ldots, m$. The number of CPOD basis functions of each class is not necessarily equal in our method, but in order to compare with the standard POD method, it is set to be equal and determined by the 97% cumulative energy ratio of the standard POD bases.

In addition to estimating absolute error statistics $\widetilde{\mathcal{E}}_K$ and $\widetilde{\mathcal{V}}_K$, we also give the estimations of relative error statistics defined as

$$\widetilde{\mathcal{E}}_K^r = \mathbb{E}\left[\frac{\|u - \widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}}{\|u\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}}\right] \tag{6.3}$$

and

$$\widetilde{\mathcal{V}}_K^r = \mathbb{V}\left[\frac{\|u - \widetilde{u}^K\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}}{\|u\|^2_{\mathcal{L}^2([0,T];\, L^2(D))}}\right]. \tag{6.4}$$

These statistics are all estimated by the MC method. Next, we consider two different strengths A, one is expanded by the trigonometric functions, and the other is hat-type functions of different heights with white noise.

6.1 Strength expanded by trigonometric functions

In this experiment, the strength A is given by

$$A(t;\, \omega) = A_0(t) + \sigma \sum_{i=1}^{N} \delta_i \left[\sin(\pi i t)\eta_i^{(1)}(\omega) + \cos(\pi i t)\eta_i^{(2)}(\omega)\right], \tag{6.5}$$

where the mean strength $A_0(t) \equiv 70$, amplification factor $\sigma = 12$, the number of expanded terms $N = 100$, $\delta_i = 1/i$ for $i = 1, 2, \ldots, N$, and $\{\eta_i^{(j)}\}_{i=1}^N$, $j = 1, 2$, are i.i.d. random variables and satisfy $\eta_i^{(j)} \sim \mathcal{N}(0, 1)$. Here, 300 samples of velocity field are used to generate the CPOD bases and train the naive Bayes pre-classifier, and the other 100 samples form the test set to estimate the error of the SROM based on the pre-classifier.

6.1.1 Generating CPOD basis functions

Figure 5 shows the clustering results of these 300 samples with modified t-gCVT method. On the left is the number of samples in each class, n_k, under different cluster numbers K. The middle is the corresponding energy defined in (3.11), which gradually decreases with the increase of K. On the right is the logarithm of eigenvalues corresponding to the first 30 CPOD basis functions in each class. The dimensions and cumulative energy ratios used in this experiment are given in Table 1. On the whole, for $K = 2$ and 3, the energy ratios of the CPOD basis functions generated by our method are higher than that of the standard POD method. It is not difficult to understand that the samples in each class are similar after clustering, so their eigenvalues decay faster, which leads to the same number of basis functions can obtain more information. That is to say, some information that is ignored by standard POD method can be captured after clustering. The contours of the first four CPOD basis functions in every class are given in Figure 6.

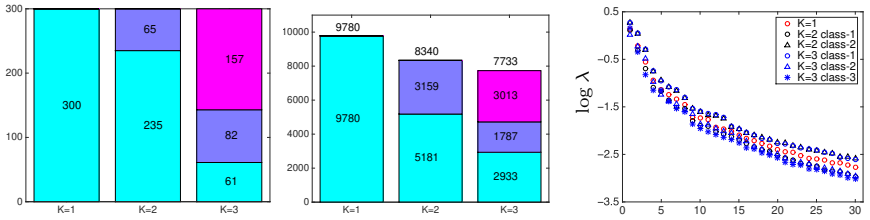

Fig. 5: Population n_k (left), energy $\widetilde{\mathcal{E}}^{\text{t-gCVT}}$ (middle) of data $\widehat{U}$, and the logarithm of eigenvalues (right) corresponding to the first 30 CPOD basis functions in each class for $K = 1, 2$ and 3

Table 1: The dimension d_k and cumulative energy ratio ν_k of CPOD basis functions in each class for $K = 1, 2$ and 3

	$K=1$	$K=2$		$K=3$		
class	-	1	2	1	2	3
d_k	16	16	16	16	16	16
ν_k	0.9704	0.9765	0.9713	0.9719	0.9768	0.9798

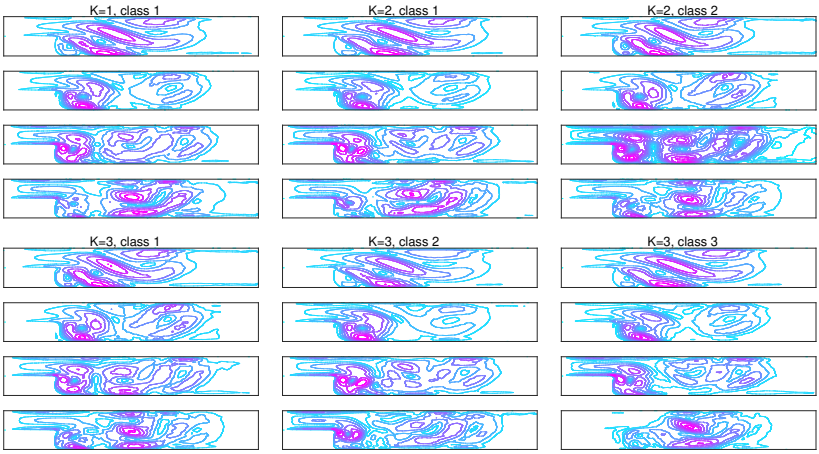

Fig. 6: Contours of the first four CPOD basis functions in each class for $K = 1, 2$ and 3

From the clustering results of modified t-gCVT, the labels of these 300 training samples are known. The errors of CPOD-based SROM that directly use the training data and their known labels are given in the Table 2, and the statistics of $L^2(D)$-norm error between finite element solution and CPOD reduced-order solution are shown in Figure 7. Clearly, when the class labels of samples are known, the CPOD-based SROM is more accurate and more stable than the standard POD-based SROM. This illustrates that it is feasible to use CPOD basis functions to improve the accuracy of the reduced-order model. Figure 8 gives the simulation results of two samples in the training set, which more intuitively shows the performance of the CPOD basis functions.

Table 2: Error estimates of CPOD-based SROM by using 300 labelled training data $\widehat{U}$

K	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$
1	0.6736	3.0114%	0.6788	0.1325%
2	0.6229	2.7493%	0.1879	0.0361%
3	0.5516	2.4526%	0.1280	0.0261%

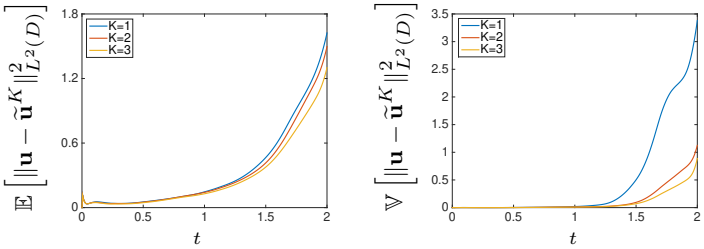

Fig. 7: Error estimates of CPOD-based SROM with different K

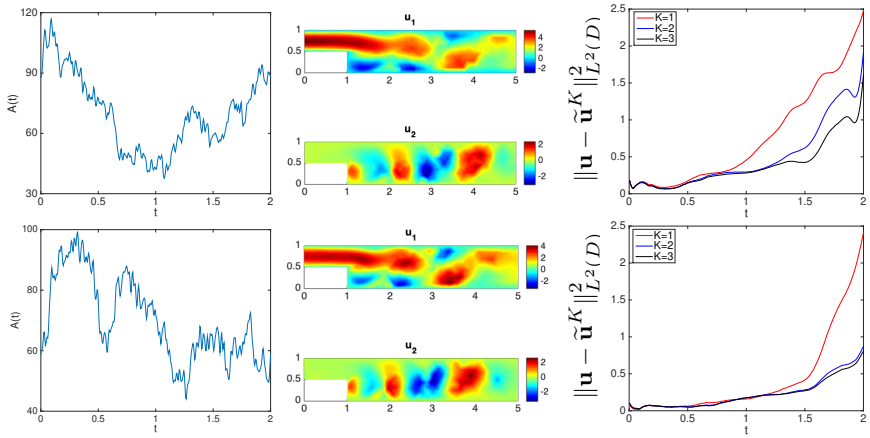

Fig. 8: Two realizations of the strength $A(t)$ in stochastic inlet velocity u_{in} (left), and their corresponding finite element solutions $\mathbf{u} = (u_1, u_2)^\top$ at time T (middle), and the errors of CPOD approximate solutions (right)

6.1.2 Simulation results of CPOD-NB based SROM

Use 300 inputs $\{\boldsymbol{\xi}_i\}_{i=1}^{300}$ associated with data set $\widehat{U}$ and the clustering results of modified t-gCVT method to train a naive Bayes pre-classifier. Here, we directly use the naive Bayes classification toolbox of MATLAB. For these 100 test data, use the pre-classifier to get their predicted labels, and use formula (3.26) to get their true labels. The resulting confusion matrices are shown in Figure 9. It can be observed that when $K = 2$, all 53 samples with the true label of 1 are predicted correctly, while 20 of the 47 samples with the true label of 2 are predicted incorrectly. In other words, the predicted labels of 80% of the test data are consistent with their true labels. Similarly, 70% of the test samples are correctly predicted for $K = 3$. As defined in (4.14), the error rates of the naive Bayes pre-classifier are 9.22% when $K = 2$ and 20.10% when $K = 3$.

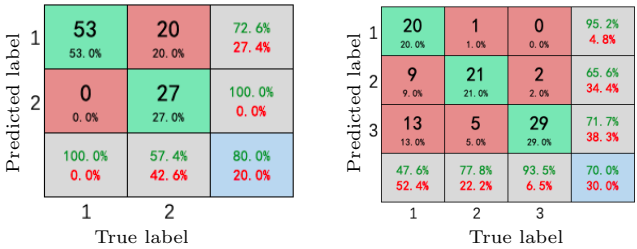

Fig. 9: Confusion matrices of test data set with 100 samples for $K = 2$ and 3

Table 3 lists the errors of the CPOD-NB based SROM estimated with the test data. The results on the left are associated with the true labels, while the results on the right are associated with predicted labels. Obviously, whether the true labels or the predicted labels are used, the accuracy of CPOD-NB based SROM is gradually improving with the increase of K, even though the misjudgment samples have an impact on the accuracy of our SROM. Figure 10 shows the errors of 4 samples in the test data. It can be seen that the reduced-order solutions calculated by our true best-matched CPOD basis functions have better accuracy than the standard POD reduced-order solution, but the errors may be larger than that of the standard POD method in the case of misjudgment.

Table 3: Error estimates of the CPOD-NB based SROM by using 100 test samples under the true labels (left) and predicted labels (right)

	True labels				Predicted labels			
K	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$
1	0.6256	2.8137%	0.4200	0.0871%	0.6256	2.8137%	0.4200	0.0871%
2	0.5062	2.2723%	0.0738	0.0160%	0.5477	2.4340%	0.0925	0.0191%
3	0.4576	2.0594%	0.0611	0.0133%	0.5038	2.2353%	0.0754	0.0156%

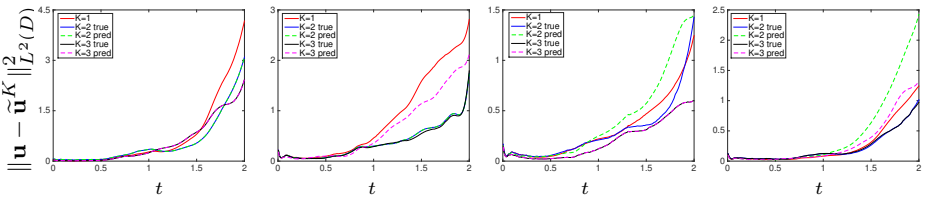

Fig. 10: The errors of the CPOD-NB approximate solutions of four samples in the test data

6.2 Hat-type strength with white noise

In this numerical experiment, the strength A takes the following form

$$A(t;\ \omega) = \sigma \frac{dW}{dt} + 60 \begin{cases} 1 + at & t \in [0, 1], \\ 1 + a(2 - t) & t \in [1, 2], \end{cases} \tag{6.6}$$

where the height parameter $a \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$, and the amplification factor of white noise $\sigma = 1.5$. The white noise $\frac{dW}{dt}$ is approximated by the piecewise constant

$$\frac{dW^m}{dt} = \frac{1}{\sqrt{Dt}} \sum_{i=0}^{m-1} \chi_i(t)\eta_i(\omega), \tag{6.7}$$

where the components of $\boldsymbol{\eta}(\omega) = [\eta_0(\omega), \ldots, \eta_{m-1}(\omega)]^\top$ are i.i.d. random variables and satisfy the standard normal distribution $\mathcal{N}(0, 1)$, and the characteristic function $\chi_i(t)$ is defined by

$$\chi_i(t) = \begin{cases} 1 & t \in [t_i, t_{i+1}), \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

Figure 11 shows the strengths A corresponding to different coefficients a when not affected by white noise.

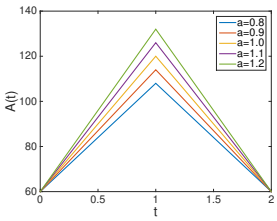

Fig. 11: Strengths A corresponding to different coefficients a ($\sigma = 0$)

Here, we take $a = 0.8, 0.9, 1.0, 1.1$ and 1.2 to generate 80 samples of velocity field respectively, and use these samples to form a data set $\widehat{U}$ for constructing the CPOD basis functions and training the naive Bayes pre-classifier. In addition, use these coefficients a to generate 20 samples respectively to form a test set for estimating the error of CPOD-NB based SROM.

6.2.1 Generating CPOD basis functions

Figure 12 shows the clustering results of these 400 training data by using the modified t-gCVT method. The dimensions and cumulative energy ratios used in this experiment are listed in Table 4. Although the energy ratios of the second class with $K = 2$ and the second and third classes with $K = 3$ are all slightly smaller than that with $K = 1$, the energy ratios of the first class

with $K = 2$ and $K = 3$ are much larger than that of the standard POD basis functions. Figure 13 shows the contours of the first four CPOD basis functions in each class for different K.

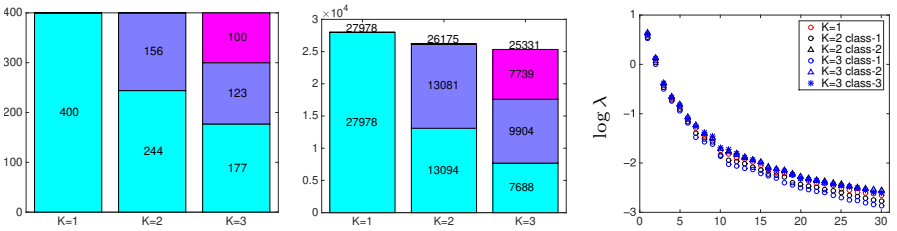

Fig. 12: Population n_k (left), energy $\widetilde{\mathcal{E}}^{\text{t-gCVT}}$ (middle) of data $\widehat{U}$, and the logarithm of eigenvalues (right) corresponding to the first 30 CPOD basis functions in each class for $K = 1, 2$ and 3

Table 4: The dimension d_k and cumulative energy ratio ν_k of CPOD basis functions in each class for $K = 1, 2$ and 3

	$K=1$	$K=2$		$K=3$		
class	-	1	2	1	2	3
d_k	11	11	11	11	11	11
ν_k	0.9714	0.9761	0.9698	0.9793	0.9710	0.9709

Table 5 gives the estimated error of the CPOD-based SROM by using 400 labelled training data. Obviously, from the perspective of expectation, the accuracy of our SROM increases with the increase of K. The variance of absolute error is also increasing, but only slightly in terms of the relative error. Figure 14 shows two samples in the training data and their errors of CPOD approximate solutions.

Table 5: Error estimates of CPOD-based SROM by using 400 labelled training data $\widehat{U}$

K	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$
1	0.6242	1.9596%	0.0454	0.0406%
2	0.5915	1.8572%	0.0515	0.0499%
3	0.5456	1.7731%	0.0519	0.0515%

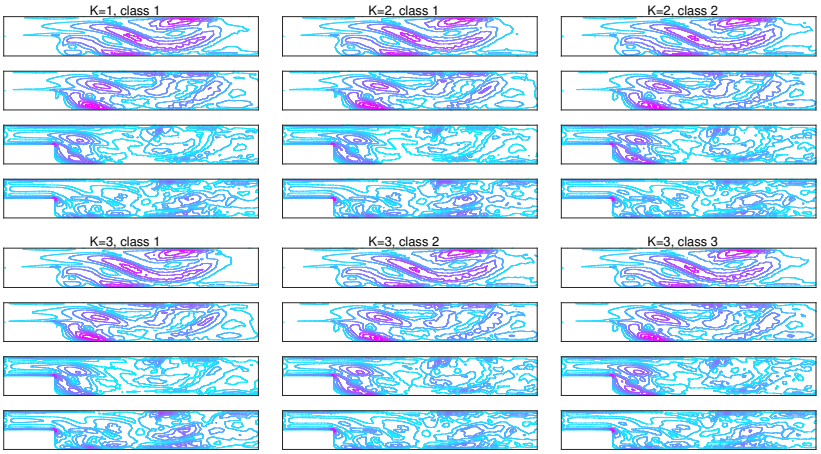

Fig. 13: Contours of the first four CPOD basis functions in each class for $K = 1, 2$ and 3

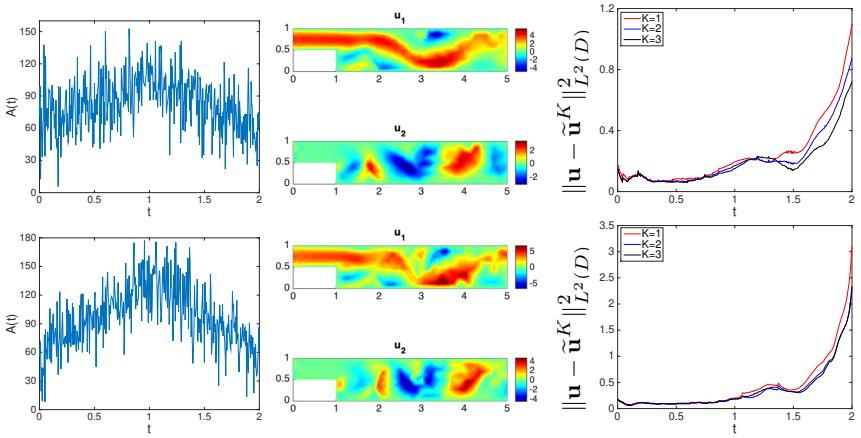

Fig. 14: Two realizations of the strength $A(t)$ in stochastic inlet velocity u_{in} (left), and their corresponding finite element solutions $\mathbf{u} = (u_1, u_2)^\top$ at time T (middle), and the errors of CPOD approximate solutions (right)

6.2.2 Simulation results of CPOD-NB based SROM

For these 100 test data, the confusion matrices are shown in Figure 15. The corresponding error rates of naive Bayes pre-classifier are 15.80% when $K = 2$ and 31.99% when $K = 3$. Although the error rate of the pre-classifier is higher for the high-dimensional data affected by white noise, our SROM can still maintain its advantages within the acceptable range. The errors of the CPOD-NB based SROM estimated by using the test data are given in Table 6. It is clearly that under the influence of misjudgment samples, our SROM still has

a significant improvement compared to the standard POD method. The errors of four samples in test set are shown in Figure 16.

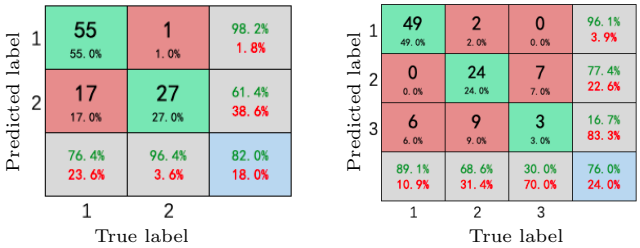

Fig. 15: Confusion matrices of test data set with 100 samples for $K = 2$ and 3

Table 6: Error estimates of the CPOD-NB based SROM by using 100 test samples under the true labels (left) and predicted labels (right)

	True labels				Predicted labels			
K	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$	$\widetilde{\mathcal{E}}_K$	$\widetilde{\mathcal{E}}_K^r$	$\widetilde{\mathcal{V}}_K$	$\widetilde{\mathcal{V}}_K^r$
1	0.6319	1.9262%	0.0478	0.0125%	0.6319	1.9262%	0.0478	0.0125%
2	0.5888	1.7472%	0.0510	0.0080%	0.6115	1.8178%	0.0555	0.0081%
3	0.5464	1.6587%	0.0502	0.0086%	0.5722	1.7240%	0.0594	0.0090%

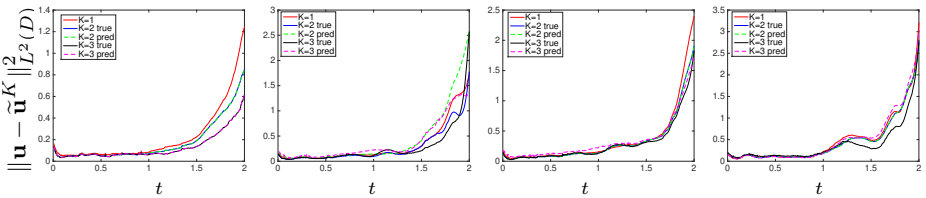

Fig. 16: The errors of the CPOD-NB approximate solutions of four samples in the test data

Compared with the results in section 6.1, it can be seen from Tables 5 and 6 that the improvement of our SROM in this experiment is relatively limited, mainly includes the following two reasons. First of all, although affected by the white noise, the strength A still shows a hat-shaped trend as a whole, so the similarity between the realizations of the velocity field is higher, thereby the resulting CPOD basis functions are less different from the standard POD basis functions. Secondly, the stronger randomness of input $\boldsymbol{\xi}$ leads to worse classification results, which increases the influence of misjudgment.

7 Conclusion

We develop a method for model reduction by combining clustering and classification. According to the mapping relationship between input and output of the system, we use the modified t-gCVT method to cluster the output samples and generate several sets of CPOD basis functions, then use the clustering results to learn the classification mechanism of input. For a given input, compared to the standard POD bases, the best-matched CPOD basis functions can reduce the model better. However, as the number of clusters increases, not only the computational complexity increase due to a large number of distance calculations, but also the error rate of the pre-classifier increases, which will affect the accuracy of our SROM. Therefore, it is necessary to study the appropriate number of clusters. In order to improve the stability of our algorithm, the classification of high-dimensional data is also a subject worth studying in the future, such as combining the state-of-the-art deep learning techniques. This paper is mainly to provide a prototype of reduced-order modelling by using statistical analysis methods, and this idea can be applied to more complex problems, such as uncertainty quantification, optimal control, etc.

Acknowledgments. The research of J. Ming is supported by ???. The research of Z. Zhang is supported by Hong Kong RGC grant projects 17300318 and 17307921. The authors would like to thank Zhipeng Yang for carefully reviewing the manuscript and interesting discussions. The authors also want to thank the editors and the reviewers of this paper for their helpful suggestions and comments.

Declarations

- **Conflict of interest/Competing interests:** The authors have no conflicts of interest to declare that are relevant to the content of this article.
- **Availability of data and materials:** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.
- **Code availability:** Custom code.

References

[1] Quarteroni, A., Rozza, G., *et al.*: Reduced Order Methods for Modeling and Computational Reduction vol. 9. Springer, Berlin (2014)

[2] Han, S., Feeny, B.: Application of proper orthogonal decomposition to structural vibration analysis. Mechanical Systems and Signal Processing **17**(5), 989–1001 (2003)

[3] Samir, K., Brahim, B., Capozucca, R., Wahab, M.A.: Damage detection in CFRP composite beams based on vibration analysis using proper orthogonal decomposition method with radial basis functions and cuckoo search

algorithm. Composite Structures **187**, 344–353 (2018)

[4] Lim, H., Wei, X., Zang, B., Vevek, U., Mariani, R., New, T., Cui, Y.: Short-time proper orthogonal decomposition of time-resolved schlieren images for transient jet screech characterization. Aerospace Science and Technology **107**, 106276 (2020)

[5] Holmes, P., Lumley, J.L., Berkooz, G., Rowley, C.W.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge university press, Cambridge (2012). https://doi.org/10.1017/CBO9780511919701

[6] Berkooz, G., Holmes, P., Lumley, J.L.: The proper orthogonal decomposition in the analysis of turbulent flows. Annual review of fluid mechanics **25**(1), 539–575 (1993)

[7] Mendez, M., Balabane, M., Buchlin, J.-M.: Multi-scale proper orthogonal decomposition of complex fluid flows. Journal of Fluid Mechanics **870**, 988–1036 (2019)

[8] Fathi, M.F., Bakhshinejad, A., Baghaie, A., Saloner, D., Sacho, R.H., Rayz, V.L., D'Souza, R.M.: Denoising and spatial resolution enhancement of 4D flow MRI using proper orthogonal decomposition and lasso regularization. Computerized Medical Imaging and Graphics **70**, 165–172 (2018)

[9] Saha, I., Sarkar, J.P., Maulik, U.: Integrated rough fuzzy clustering for categorical data analysis. Fuzzy Sets and Systems **361**, 1–32 (2019)

[10] Agresti, A.: An Introduction to Categorical Data Analysis. John Wiley & Sons, New York, NY (2018)

[11] Yee, T.W., *et al.*: The VGAM package for categorical data analysis. Journal of Statistical Software **32**(10), 1–34 (2010)

[12] Savalei, V.: Improving fit indices in structural equation modeling with categorical data. Multivariate behavioral research, 1–18 (2020)

[13] Xu, R., Wunsch, D.: Clustering vol. 10. John Wiley & Sons, Piscataway, NJ, USA (2008)

[14] Rokach, L., Maimon, O.: Clustering methods. In: Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, Boston, MA (2005). https://doi.org/10.1007/0-387-25465-X_15

[15] Clifford, H.T., Stephenson, W., *et al.*: An Introduction to Numerical Classification vol. 240. Academic press, New York (1975)

[16] Du, Q., Faber, V., Gunzburger, M.: Centroidal Voronoi tessellations:

Applications and algorithms. SIAM review **41**(4), 637–676 (1999)

[17] Burkardt, J., Gunzburger, M., Lee, H.-C.: Centroidal Voronoi tessellation-based reduced-order modeling of complex systems. SIAM Journal on Scientific Computing **28**(2), 459–484 (2006)

[18] Burkardt, J., Gunzburger, M., Lee, H.C.: POD and CVT-based reduced-order modeling of Navier-Stokes flows. Computer Methods in Applied Mechanics & Engineering **196**(1-3), 337–355 (2006)

[19] Du, Q., Gunzburger, M.D.: Centroidal Voronoi tessellation based proper orthogonal decomposition analysis. In: Control and Estimation of Distributed Parameter Systems. International Series of Numerical Mathematics, vol. 143, pp. 137–150. Birkhauser Verlag AG, Basel, Switzerland (2003)

[20] Kaiser, E., Noack, B.R., Cordier, L., Spohn, A., Segond, M., Abel, M., Daviller, G., Östh, J., Krajnović, S., Niven, R.K.: Cluster-based reduced-order modelling of a mixing layer. Journal of Fluid Mechanics **754**, 365–414 (2014)

[21] Lee, H.-C., Lee, S.-W., Piao, G.-R.: Reduced-order modeling of Burgers equations based on centroidal Voronoi tessellation. Int. J. Numer. Anal. Model **4**(3-4), 559–583 (2007)

[22] Du, Q., Gunzburger, M., Ju, L.: Advances in studies and applications of centroidal Voronoi tessellations. Numerical Mathematics: Theory, Methods and Applications **3**(2), 119–142 (2010)

[23] Bright, I., Lin, G., Kutz, J.N.: Compressive sensing based machine learning strategy for characterizing the flow around a cylinder with limited pressure measurements. Physics of Fluids **25**(12), 127102 (2013)

[24] Bright, I., Lin, G., Kutz, J.N.: Classification of spatiotemporal data via asynchronous sparse sampling Application to flow around a cylinder. Multiscale Modeling & Simulation **14**(2), 823–838 (2016)

[25] Brunton, S.L., Tu, J.H., Bright, I., Kutz, J.N.: Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems. SIAM Journal on Applied Dynamical Systems **13**(4), 1716–1732 (2014)

[26] Brunton, S.L., Noack, B.R., Koumoutsakos, P.: Machine learning for fluid mechanics. Annual Review of Fluid Mechanics **52**, 477–508 (2020)

[27] Kramer, B., Grover, P., Boufounos, P., Nabi, S., Benosman, M.: Sparse sensing and DMD-based identification of flow regimes and bifurcations

in complex flows. SIAM Journal on Applied Dynamical Systems **16**(2), 1164–1196 (2017)

[28] San, O., Maulik, R., Ahmed, M.: An artificial neural network framework for reduced order modeling of transient flows. Communications in Nonlinear Science and Numerical Simulation **77**, 271–287 (2019)

[29] Rish, I., *et al.*: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, pp. 41–46 (2001)

[30] Murphy, K.P., et al.: Naive bayes classifiers. University of British Columbia **18**(60) (2006)

[31] Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7), 1145–1159 (1997). https://doi.org/10.1016/S0031-3203(96)00142-2

[32] Quinlan, J.R.: Simplifying decision trees. International journal of man-machine studies **27**(3), 221–234 (1987)

[33] Wang, L.: Support Vector Machines: Theory and Applications vol. 177. Springer, New York (2005)

[34] Hamerly, G., Drake, J.: Accelerating Lloyd's algorithm for k-means clustering. In: Partitional Clustering Algorithms, pp. 41–78. Springer, Berlin (2015)

[35] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer, New York (2013)

[36] Gunzburger, M., Ming, J.: Optimal control of stochastic flow over a backward-facing step using reduced-order modeling. SIAM Journal on Scientific Computing **33**(5), 2641–2663 (2011)