Combustion and Flame 187 (2018) 30–41

Contents lists available at ScienceDirect

Combustion and Flame

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

Principal component analysis coupled with nonlinear regression for chemistry reduction

Mohammad Rafi Malik[a,c,*], Benjamin J. Isaac[b], Axel Coussement[a,c], Philip J. Smith[b], Alessandro Parente[a,c,*]

[a] Université Libre de Bruxelles, Ecole Polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium
[b] Department of Chemical Engineering, University of Utah, Salt Lake City, UT 84112, USA
[c] Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Bruxelles, Belgium

ARTICLE INFO

Article history:
Received 7 March 2017
Revised 14 April 2017
Accepted 21 August 2017
Available online 21 September 2017

Keywords:
Combustion
Nonlinear regression
Local regression
Low-dimensional manifolds
Principal component analysis

ABSTRACT

Large kinetic mechanisms are required in order to accurately model combustion systems. If no parameterization of the thermo-chemical state-space is used, solution of the species transport equations can become computationally prohibitive as the resulting system involves a wide range of time and length scales. Parameterization of the thermo-chemical state-space with an a priori prescription of the dimension of the underlying manifold would lead to a reduced yet accurate description. To this end, the potential offered by Principal Component Analysis (PCA) in identifying low-dimensional manifolds is very appealing. The present work seeks to advance the understanding and application of the PC-transport approach by analyzing the ability to parameterize the thermo-chemical state with the PCA basis using nonlinear regression. In order to demonstrate the accuracy of the method within a numerical solver, unsteady perfectly stirred reactor (PSR) calculations are shown using the PC-transport approach. The PSR analysis extends previous investigations to more complex fuels (methane and propane), showing the ability of the approach to deal with relatively large kinetic mechanisms. The ability to achieve highly accurate mapping through Gaussian Process based nonlinear regression is also shown. In addition, a novel method based on local regression of the PC source terms is also investigated which leads to improved results.

1. Introduction

The numerical modeling of turbulent combustion is a very challenging task as it combines the complex phenomena of turbulence and chemical reactions. This study becomes even more challenging when large detailed kinetic mechanisms are used in order to understand some special features such as pollutant formation. A detailed combustion mechanism for a simple fuel such as methane involves 53 species and 325 chemical reactions [1]. Moreover, the number of species and reactions increases with increasing fuel complexity. The coupling of the kinetic equations with the set of Navier–Stokes equations results in a problem that is too complex to be solved by the current computational means. In a CFD calculation, the number of species tracked impacts the memory usage and CPU time. It is thus important to minimize this number by the use of a simpler but representative set of variables. Therefore, there is a need for methods allowing to parameterize efficiently the thermo-chemical state of a reacting system with a reduced number of optimal reaction variables. Among those, Principal Component Analysis (PCA) appears as an ideal candidate to fulfill the purpose [2–8]. PCA offers the possibility of automatically reducing the dimensionality of data sets consisting of a large number of correlated variables, while retaining most of the variation present in the original data. After reduction, the new set of variables, called principal components (PCs), are orthogonal, uncorrelated and linear combinations of the original variables. By retaining the PCs containing most of the variance and transporting them in a numerical simulation, the dimensionality of the system can be highly reduced. Another advantage of PCA resides in the fact that the PCs can be obtained through data sets based on simple systems (such as canonical reactors) and then applied to a similar, more complex system [9]. A methodology based on PCA was proposed [5] for the identification of the controlling dynamics in reacting systems and for the consistent reduction of very large kinetic mechanisms. Sutherland and Parente [8] proposed a combustion model based on the concepts from PCA (PC-score approach). They derived transport equations for the principal components (PCs), and proposed a model where the state-space variables are constructed directly

* Corresponding authors.
E-mail addresses: rafi.malik@ulb.ac.be (M.R. Malik), Alessandro.Parente@ulb.ac.be (A. Parente).

http://dx.doi.org/10.1016/j.combustflame.2017.08.012
0010-2180/

from the PCs. The PCA-based modeling approach was enhanced [3,10,11] by combining PCA with nonlinear regression techniques, allowing a nonlinear mapping of the thermo-chemical state and the corresponding source terms onto the basis identified by the principal components. As a result, the nonlinear nature of chemical manifolds is better captured, thus, maximizing the potential size reduction provided by the method. Isaac et al. [4] and Echekki and Mirgolbabaei [2] provided the first a posteriori studies on the use of the PC-score approach. In particular, Isaac et al. showed in [4] the potential of PC-transport based combustion models coupled with nonlinear regression techniques. The model was tested on an unsteady calculation of a perfectly stirred reactor (PSR) burning syngas. The authors showed that Gaussian Process Regression (GPR) technique produced the most accurate reconstruction, showing remarkable accuracy for the prediction of temperature and major and minor species with 2 transported variables instead of 11. The approach was also tested for the first time within a CFD solver.

The present work seeks to advance the understanding and application of the PC-transport approach by applying this method to more complex fuels such as methane and propane. First, 0-D simulation of a PSR is used to generate the database for model training. Then, the solution of a steady and unsteady PSR calculation using the PC-transport approach for large kinetic mechanisms is compared with the full solution. Next, the PC-transport approach is coupled with nonlinear regression (PC-GPR) in order to increase the size reduction potential of PCA. Finally, the first study on an enhancement of the classical PC-transport approach by the use of local nonlinear regression (PC-L-GPR) is also shown. It should be pointed out that the objective of the present work was to demonstrate the applicability of GPR regression for accurate source term regression. To this purpose, the choice of a PSR is quite obvious as it allows to focus on such an aspect without the influence of transport processes.

2. Principal component analysis

Principal Component Analysis [12] is a useful statistical technique that has found application in combustion for its ability of identifying low-dimensional manifolds. In high dimension data sets, where graphical representation is not possible, PCA can be a powerful tool as it identifies correlations and patterns in a data set. Once these patterns have been identified, the data set can be compressed by reducing the number of dimensions without much loss of information. PCA analyzes the covariance between variables in a data set and identifies a linear representation of the system through orthogonal vectors, each one having a significance proportional to its eigenvalue.

In order to perform principal component analysis, a data-set $\mathbf{X}(n \times Q)$ consisting of n observations of Q independent variables is needed. Then, the data must be centered (by subtracting its mean) and scaled (using an appropriate scaling method): centering is used to convert observations into fluctuations over the mean, while scaling is done in order to compare the data evenly (if they have different units or order of magnitudes):

$$\mathbf{X_{SC}} = (\mathbf{X} - \overline{\mathbf{X}})\mathbf{D^{-1}} \tag{1}$$

where $\overline{\mathbf{X}}$ is $(n \times Q)$ matrix containing the mean of each variable and $\mathbf{D}$ is a $(n \times Q)$ matrix containing the scaling factor of each variable. Several scaling methods can be found in the literature: *auto scaling, range scaling, pareto scaling, variable stability scaling* and *level scaling* [6].

Then, one can compute the covariance matrix $\mathbf{S}$ defined as (the notation $\mathbf{X}$ will be used in the following instead of $\mathbf{X_{SC}}$ for the sake of simplicity):

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{X^T X}$$

The diagonal elements of $\mathbf{S}$ represent the variance of each variable, while the off-diagonal values show the covariance between two variables. Since $\mathbf{S}$ is a square matrix (of size $(Q \times Q)$), an eigenvalue decomposition can be performed yielding the eigenvectors and eigenvalues of the system:

$$\mathbf{S} = \mathbf{ALA^T}$$

where $\mathbf{A}$ $(Q \times Q)$ and $\mathbf{L}$ $(Q \times Q)$ are respectively the eigenvectors of $\mathbf{S}$ (also called principal components, PCs) and the eigenvalues of $\mathbf{S}$, in decreasing order. The eigenvectors matrix $\mathbf{A}$, also called the basis matrix, is used to obtain the principal component scores, $\mathbf{Z}(n \times Q)$, by projecting the original data set $\mathbf{X}$ on that basis:

$$\mathbf{Z} = \mathbf{XA} \tag{2}$$

Eq. (2) indicates that the original data set can be uniquely recovered using the PCs and their scores:

$$\mathbf{X} = \mathbf{ZA^{-1}}$$

where $\mathbf{A^{-1}} = \mathbf{A^T}$. Then, using a subset of $\mathbf{A}$ by retaining only q PCs (with $q < Q$), noted $\mathbf{A_q}$, an approximation of $\mathbf{X}$ based on the first q eigenvectors ($\mathbf{X_q}$) is obtained:

$$\mathbf{X} \approxeq \mathbf{X_q} = \mathbf{Z_q A_q^T}$$

where $\mathbf{X_q}$ is the approximation of $\mathbf{X}$ based on the first q eigenvectors of Q, and $\mathbf{Z_q}$ is the $(n \times q)$ matrix of the principal component scores. In the PC analysis, the largest eigenvalues correspond to the first columns of $\mathbf{A}$. This means the largest amount of variance in the original variables is described by the first PCs. Thus, the truncation is made on the last eigenvectors (corresponding to the smallest eigenvalues). By removing the last PCs, the dimension of the system is reduced while retaining most of the variation in the system.

2.1. PC-score approach

In the work of Sutherland and Parente [8], a model based on transport equations for the PCs is proposed derived from the general species transport equation:

$$\frac{\partial}{\partial \mathbf{t}}(\rho \mathbf{Y_k}) + \nabla(\rho \mathbf{\tilde{u} Y_k}) = \nabla(\rho \mathbf{D_k} \nabla \mathbf{Y_k}) + \mathbf{R_k} \qquad k = 1, ..., n_s \tag{3}$$

where Y_k is the mass fraction of species k and R_k is its corresponding source term (with n_s the total number of species in the system), D_k the diffusion coefficient for species k, ρ the density and $\tilde{u}$ the velocity vector. Transport equations for the PC scores ($\mathbf{Z}$) can be formulated from Eq. (3) given the basis matrix $\mathbf{A}$ and the scaling factors d_k:

$$\frac{\partial}{\partial \mathbf{t}}(\rho \mathbf{z}) + \nabla(\rho \mathbf{\tilde{u} z}) = \nabla(\rho \mathbf{D_z} \nabla \mathbf{z}) + \mathbf{s_z} \tag{4}$$

$$\mathbf{s_z} = \sum_{\mathbf{k=1}}^{\mathbf{Q}} \frac{\mathbf{R_k}}{\mathbf{d_k}} \mathbf{A_{kq}} \tag{5}$$

where $\mathbf{z} = \mathbf{Z_i^t}$ represents an individual score realization. One of the major weaknesses of classic PCA is that a multi-linear model is used to approximate a highly nonlinear manifold. The nonlinearity of chemical manifolds can be attributed to the high nonlinearity of chemical source terms (Arrhenius). This can be visualized in Fig. 1, showing the first principal component source term s_{z1}, as a function of the first two principal components for the propane case.

In the present work, PCA is used to identify the most appropriate basis to parameterize the empirical low-dimensional manifolds and define transport equations in the new space (see Eq. 4 and 5). Then, both the state space and the source terms are nonlinearly regressed onto the new basis using several approaches, described in Section 3. The nonlinear regression of the chemical state