Combustion and Flame 160 (2013) 340–350

Contents lists available at SciVerse ScienceDirect

Combustion and Flame

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

Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity

Alessandro Parente [a],*, James C. Sutherland [b]

[a] Service d'Aéro-Thermo-Mécanique, Université, Libre de Bruxelles, Bruxelles, Belgium
[b] Department of Chemical Engineering, University of Utah, Salt Lake City, UT 84112, USA

ARTICLE INFO

Article history:
Received 7 February 2012
Received in revised form 10 July 2012
Accepted 29 September 2012
Available online 7 November 2012

Keywords:
Combustion
Low-dimensional manifolds
Outliers
Pre-processing
Principal component analysis
Scaling

ABSTRACT

Principal component analysis has demonstrated promise in its ability to identify low-dimensional chemical manifolds in turbulent reacting systems by providing a basis for the *a priori* parameterization of such systems based on a reduced number of parameterizing variables. Previous studies on PCA have only mentioned the importance of data pre-processing and scaling on the PCA analysis, without detailed consideration. This paper assesses the influence of data-preprocessing techniques on the size-reduction process accomplished through PCA. In particular, a methodology is proposed to identify and remove outlier observations from the datasets on which PCA is performed. Moreover, the effect of centering and scaling techniques on the PCA manifold is assessed and discussed in detail, to investigate how different scalings affect the size of the manifold and the accuracy in the reconstruction of the state-space. Finally, the sensitivity of the chemical manifold to flow characteristics is considered, to investigate its invariance with respect to the Reynolds number. Several high-fidelity experimental datasets from the TNF workshop database are considered in the present work to demonstrate the effectiveness of the proposed methodologies.

1. Introduction

Recently, principal component analysis (PCA) was introduced as a method of identifying manifolds in turbulent combustion [1]. PCA has also been used by others to analyze combustion data [2–4], but for different purposes – see [1] for a discussion. The merits of PCA in the context of modeling turbulent reacting flows have been demonstrated for identifying low-dimensional manifolds underlying the thermo-chemical state [1,5] and toward the development of PCA-based combustion models [6,7]. A particularly noteworthy feature of PCA-based models is the possibility of obtaining low-dimensional parameterizations satisfying well-defined error bounds. Previous studies on PCA [1,5] have mentioned the importance of pre-processing data prior to applying PCA, but the effects of pre-processing strategies have not been assessed in detail. In particular, the effect of potential outlier observations as well as the role of centering and scaling on the principal component structure has not been addressed. The objective of the present paper is to review the PCA procedure and highlight the role of the available pre-processing techniques on the robustness of PCA and its ability to identify a low-dimensional representation of a thermo-chemical manifold. The sensitivity of PCA to modifications of the database from which the low-dimensional basis is extracted is also considered, to investigate the universality of the PCA method.

Section 2 provides a review of PCA as well as a discussion on outlier removal (2.1), data centering and scaling (2.2), and dimension reduction (2.3). Section 3 applies PCA to several experimental datasets from the Sandia non-premixed flame datasets to illustrate the effect of pre-processing and scaling on the PCA reduction. Finally, the invariance of the chemical manifold with respect to the Reynolds number is demonstrated for a set of piloted flames at a range of Reynolds numbers.

2. Principal component analysis

Principal component analysis (PCA) [8,9] provides a rigorous mathematical formalism for the identification of the most active directions in multivariate datasets. PCA identifies correlations among the variables defining the state space. As a result, a new coordinate system is identified in the directions of maximal data variance, which allows less important dimensions to be eliminated while maintaining the primary structure of the original data. Details of the PCA reduction have been already provided [1]. Here, the PCA concept will be reviewed briefly whereas the impact of

* Corresponding author. Address: Avenue F.D. Roosevelt 50, 1050 Bruxelles, Belgium. Fax: +32 2 650 27 10.
E-mail address: Alessandro.Parente@ulb.ac.be (A. Parente).

0010-2180/$ - see front matter © 2012 The Combustion Institute. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.combustflame.2012.09.016

pre-processing and post-processing on PCA results will be discussed in detail.

In PCA, n observations of Q variables are assigned to an $(n \times Q)$ matrix $\mathbf{X}$ whose rows represent individual observations of all Q variables $\boldsymbol{x}$. For the combustion applications considered in this paper, the Q columns in $\mathbf{X}$ are taken to be the temperature and species mass fractions.[1] PCA projects $\boldsymbol{x}$ onto a rotated basis obtained from the eigenvalue decomposition of the $(Q \times Q)$ covariance matrix,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{A}\mathbf{L}\mathbf{A}^T, \tag{1}$$

where $\mathbf{A}$ and $\mathbf{L}$ are the eigenvectors and eigenvalues of $\mathbf{S}$. The rotated basis, defined by the eigenvectors $\mathbf{A}$, may be truncated to retain the most energetic directions (those columns of $\mathbf{A}$ associated with the largest eigenvalues of $\mathbf{L}$), providing the non-square matrix $\mathbf{A}_q$ on which the original data are projected to obtain the principal components (PC), $\mathbf{Z}_q$,

$$\mathbf{Z}_q = \mathbf{X}\mathbf{A}_q. \tag{2}$$

Eq. (2) can be inverted to obtain an approximate reconstruction of the original $(n \times Q)$ dimensional sample:

$$\mathbf{X}_q = \mathbf{Z}_q\mathbf{A}_q^T. \tag{3}$$

Eq. (3) is a linear reconstruction. The intrinsic linearity of the PCA approach represents a major possible drawback of the technique to deal with strongly non-linear processes such as combustion. However, this limitation can be partially overcome using local PCA [1,10]. Alternatively, non-linear reconstructions can provide more accurate mappings from $\mathbf{Z}_q$ to $\mathbf{X}_q$ [7]. The PCA reduction process is represented schematically in Fig. 1.

Several procedures are required prior to performing the PCA reduction process (Fig. 1):

1. *Outlier removal.* Experimental datasets usually contain a few unusual observations which can strongly affect the data covariance structure and, therefore, the structure of the principal components. If we refer to a one-dimensional problem, the outliers can be classified as those observations which are either very large or very small with respect to the others. In high dimensions, there can be outliers that do not appear as outlying observations when considering each dimension separately and, therefore, they will not be detected using univariate criteria. Thus, a multivariate approach must be pursued. PCA itself represents an ideal tool for the identification and removal of outlier observations.
2. *Centering and scaling*. Data are usually *centered* and *scaled* before PCA is carried out. Centering represents all observations as fluctuations, leaving only the relevant variation for analysis. Scaling is a crucial operation when analyzing the thermochemical state of a reacting system since temperature and species concentrations have different units and vary over different scales. The choice of scaling significantly affects the subsequent PCA analysis: different scalings allow to emphasize correlations among different groups of state variables, providing an effective tool for targeting the PCA analysis on the variables which are most relevant for an investigated application.

Section 2.1 presents a technique to identify outliers, while Section 2.2 addresses centering and scaling.

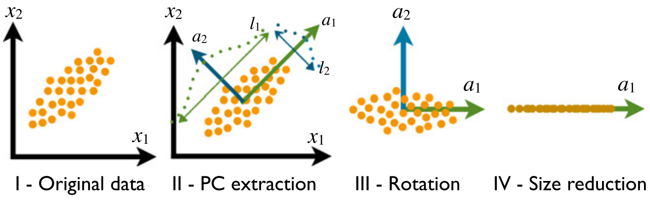

Fig. 1. PCA reduction process.

2.1. Outlier detection and removal with PCA

The usual procedure for outlier detection in multivariate data analysis is to measure the distance of each realization i of the Q observed variables, from the data center, using the so called Mahalanobis distance:

$$D_M = (\mathbf{X} - \overline{\mathbf{X}})^T\mathbf{S}^{-1}(\mathbf{X} - \overline{\mathbf{X}}), \tag{4}$$

where $\overline{\mathbf{X}}$ is a matrix containing the average values, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, of the original variables. The observations associated with large values of D_M are classified as outliers and then discarded. The Mahalanobis distance can be related to the principal components: it can be shown, in fact, that the sum of squares of the PC, standardized by the eigenvalue size, equals the Mahalanobis distance for observation i:

$$\sum_{k=1}^{Q}\frac{z_{ik}^2}{l_k} = \frac{z_{i1}^2}{l_1} + \frac{z_{i2}^2}{l_2} + \cdots + \frac{z_{iQ}^2}{l_Q} = D_{M,i}. \tag{5}$$

This realization can be exploited for building a robust methodology based on PCA for outlier identification and removal. As mentioned previously, the first few principal components have large variances and explain most of the variation in $\mathbf{X}$. Therefore, those components are strongly affected by variables with relatively large variances and covariances. Consequently, the observations that are outliers with respect to the first few components usually correspond to outliers on one or more of the original variables. On the other hand, the last few principal components represent linear functions of the original variables with minimal variance. These components are sensitive to the observations that are inconsistent with the covariance structure of the data but are not outliers with respect to the original individual variables. Based on the above considerations, the following detection scheme can be proposed, as suggested by [11]:

1. *Multivariate trimming.* A fraction γ of the data points characterized by the largest value of D_M are classified as outliers and removed. $\overline{\mathbf{X}}$ and $\mathbf{S}$ are then computed from the remaining observations. The trimming process can be iterated to ensure that $\overline{\mathbf{X}}$ and $\mathbf{S}$ are resistant to outliers.
2. *Principal components classifier*. The classifier consists of two functions, one from the major, $\sum_{k=1}^{q}\frac{z_{ik}^2}{l_k}$, and one from the minor principal component, $\sum_{k=Q-r+1}^{Q}\frac{z_{ik}^2}{l_k}$. The first function can easily detect observations with large values on some of the original variables; in addition, the second function helps detect the observations that do not conform to the covariance structure of the sample. The number of major components, q, is determined by retaining the minimum number of PC required to account for at least 50% of the original data variance, while r is chosen so that the minor components used for the definition of the classifiers are those whose variance is less than $0.2 \cdot \bar{l}$, where $\bar{l}$ is the average value of the eigenvalues of $\mathbf{S}$. This ensures that the selected minor components account for a very marginal variance and they only represent linear relations among the variables. Based on the above definition, an observation $\mathbf{X}_i$ is classified as an outlier if:

[1] Formally, pressure should also be included, but for low mach number flows in open domains, it is safely neglected.