A Comparative Study of High-Order Singular Value Decomposition and Principal Component Analysis for Combustion Data Analysis

Isacco Faglioni, Soledad Le Clainche, Laura Saveedra Lago, Alessandro Parente

Universidad Politécnica de Madrid

1) Novelty Statement

The study of tensorial techniques for combustion data analysis and reduced order modeling is a rapidly evolving field. We provide the first explicit mathematical interpretation linking PCA and its tensorial generalization, HOSVD, within the context of combustion data analysis. Our results, validated on DNS data of a hydrogen flame, bridges the gap between traditional matrix-based and higher-order reduced-order models in turbulent reacting flows

2) Significance Statement

Currently there is no explicit link between reduced order models techniques based on SVD such as Principal Component Analysis (PCA) and their tensorial counterpart. This work interprets the various components of these techniques when applied specifically to a chemical reactive flow and proves the mathematical equivalence of some of their elements. The equivalence is checked numerically on a direct simulation of a hydrogen flame. On the same dataset the improvements compression-wise of HOSVD are measured.

3) Author Contribution

- I. F. designed research, performed analysis, wrote the paper ...
- S. lC. M. supervision
- L. S. L. supervision
- A. P. designed research, supervision

4) Preferred Presentation Format

The authors request **oral presentation format** for the following reasons: a) the mathematical framework combining HOSVD, PCA equivalence, and core tensor analysis requires detailed explanation that is easier to present orally; b) the visualization of high-dimensional tensor decompositions is easier to describe juxtaposed with the presenter's oral explanation.

5) Colloquium/Colloquia Designation and Keywords

Colloquium choices (order of preference): a) Modeling Approaches; b) Numerical Techniques

Keywords: Tensor decomposition; High-order singular value decomposition; Principal component analysis; Reduced-order modeling;

A Comparative Study of High-Order Singular Value Decomposition and Principal Component Analysis for Combustion Data Analysis

Isacco Faglioni[*,a,b],Soledad Le Clainche[a], Laura Saveedra Lago[a], Alessandro Parente[b,c,d]

[a] *Universidad Politécnica de Madrid*
[b] *Aero-Thermo-Mechanics Department, Université Libre de Bruxelles (ULB), Brussels, Belgium*
[c] *Brussels Institute for Thermal-Fluid Systems and Clean Energy (BRITE), ULB and VUB, Brussels, Belgium*
[d] *WEL Research Institute, Wavre, Belgium*

Abstract

Tensorial methods for dimensionality reduction and reduced order modeling in combustion are gaining popularity. An important step for this research direction is to understand how these methods relate with respect to established reduced order model techniques used by the community. This study presents a comparison between Principal Component Analysis (PCA) and High Order Singular Value Decomposition (HOSVD) from a theoretical and practical point of view. A mathematical proof of the equivalence for the corresponding components of the two methods is provided, proving the complete equivalence of PCA loadings and generalized singular value decay. An explanation of the physical meaning of the HOSVD components is provided, showing how the approach can be used to interpret reacting flow data. The theoretical results are checked numerically on a direct numerical simulation of a jet hydrogen flame. Finally, a comparison between HOSVD and PCA on the same dataset shows how the former can lead to compression ratios three orders of magnitude higher with comparable errors.

Keywords: Tensor decomposition; High-order singular value decomposition; Principal component analysis; Reduced-order modeling

Corresponding author *

1. Introduction

Experimental measurements and simulation of combustion phenomena produce data that vary in both space and time. The latter can be represented in tensor form, which contains components associated with space, time and thermodynamic variables. Extracting information from these high-dimensional datasets requires dimensionality reduction techniques that can identify dominant patterns and reconstruct with prescribed precision the original data from a compressed state. In recent years, Principal Component Analysis (PCA) has been successfully employed to project high-dimensional data onto a lower-dimensional orthogonal subspace of thermodynamical variables [1]. However, PCA is by construction limited to matrix data, requiring the flattening of multi-dimensional spatiotemporal combustion data into two-dimensional arrays, losing the natural structure and correlations that could be otherwise preserved in the tensor structure. High Order Singular Value Decomposition (HOSVD) [2] enables the simultaneous analysis of multiple coupled variables while preserving their spatiotemporal structure. Therefore, it was recently assessed in a priori analysis of Direct Numerical Simulation (DNS) data [3, 4] with the main focus on compressing and cleaning data. To the Authors knowledge, no investigation was yet devoted to benchmark HOSVD and PCA for the analysis of combustion data, particularly from the physical interpretation of the associated results. In this work, we establish a connection between PCA and HOSVD by providing a mathematical proof of their equivalence under single-mode truncation, demonstrating the correspondence between the principal components loadings and eigenvalues in PCA and those obtained from HOSVD. The mathematical results are verified numerically relying DNS data from the BLASTNET database [5, 6]. For the same dataset, HOSVD and PCA compression capabilities are benchmarked, confirming the superior ability of HOSVD to compress information when analyzing combustion data.

2. High Order Singular Value Decomposition for Combustion

Consider a spatiotemporal combustion dataset represented as a fourth-order tensor $\mathcal{X} \in \mathbb{R}^{I_x \times I_y \times I_{\text{chem}} \times I_t}$, where I_x and I_y denote

the two spatial dimensions (x and y), I_{chem} represents the number of thermodynamical variables, and I_t corresponds to time steps. The HOSVD decomposes this tensor into a core tensor and orthogonal factor matrices along each mode:

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \times_4 \mathbf{U}^{(4)} \quad (1)$$

where $\mathcal{G} \in \mathbb{R}^{I_x \times I_y \times I_{\text{chem}} \times I_t}$ is the core tensor containing the interaction coefficients between modes, $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ are the orthogonal factor matrices for mode n, and $\times_n$ denotes the mode-n product operation. Specifically, $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ capture spatial structures in x and y directions, $\mathbf{U}^{(3)}$ identifies thermochemical correlations, and $\mathbf{U}^{(4)}$ represents temporal evolution. The factor matrices are obtained by computing the SVD of the mode-n unfolding (matricization) of $\mathcal{X}$. The mode-n unfolding $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \cdots I_{n-1} I_{n+1} \cdots I_4)}$ rearranges the tensor into a matrix where mode n forms the rows and all other modes are collapsed into the columns. The SVD of each unfolding yields:

$$\mathbf{X}_{(n)} = \mathbf{U}^{(n)} \mathbf{\Sigma}^{(n)} (\mathbf{V}^{(n)})^T \quad (2)$$

where $\mathbf{\Sigma}^{(n)}$ contains the generalized singular values (mode-n singular values) ordered by decreasing magnitude. These matrices contains therefore the information that in previous formulations were associated to principal components.

The core tensor is then computed as:

$$\mathcal{G} = \mathcal{X} \times_i (\mathbf{U}^{(i)})^T \quad (3)$$

2.1. Core tensor analysis

The singular values of the Tucker decomposition can be extracted by computing the Frobenius norm of the core tensor along each mode. Specifically, for a mode α, the effective singular value associated with the i-th component is given by:

$$\sigma_\alpha(i) = \|\mathcal{G}_{i,\ldots}\|_F \quad (4)$$

where the Frobenius norm is computed over all dimensions except α. This provides a measure of the importance of each mode component in the decomposition and we will prove that it is precisely the singular value found by PCA in that dimensions. For reactive flows, this means that one can get a measure of which dimension is most suitable to be reduced. The decay of this values brings information about how much information a dimension of the system contains. In reactive flow analysis, the "slope" of the singular value decay yields physical information about the system: if we see really steep decay in spatial structure, then the system is well approximated by looking at its dynamic on a low dimensional spatial manifold.

Beyond singular values, the core tensor contains additional information about how modes along different axes of the tensor interact between each other. For a four-dimensional core tensor $\mathcal{G} \in \mathbb{R}^{R_x \times R_y \times R_{\text{chem}} \times R_t}$, the coupling strength between mode i along dimension α and the remaining dimensions can be computed as:

$$c_\alpha(i) = \sum_{j_1, j_2, j_3} |\mathcal{G}_{i,j_1,j_2,j_3}| \quad (5)$$

where the sum is taken over all indices of the dimensions excluding α. Modes with high coupling strength are those that interact significantly with modes along other dimensions. This means that the method is potentially suitable to study coupling between for example chemistry and temporal variations, or chemistry and spatial structures.

2.2. Dataset

The dataset used for this study is a DNS of a diluted, partially premixed hydrogen-air lifted flame [6] from the BLASTNET database [7]. The simulation resolves eight chemical species (H, H_2, O, O_2, OH, H_2O, HO_2, H_2O_2) on a spatial grid of 1600×2000 points across 200 temporal snapshots. To reduce computational cost of the algorithm, data was subsampled taking 1 every fifth point in each spatial dimension. Moreover, the procedure has been tested over multiple Reynolds numbers, producing results which for the scope of this work are completely equivalent. The following discussion is performed on the dataset corresponding to Re = 8000.

2.3. Preprocessing

Since chemical reactions occur based on molar quantities rather than mass, the initial step in data preparation is to convert mass fractions into molar fractions using the species' respective molar masses. One of the critical aspects of applying linear methods to combustion data is that species concentrations span several orders of magnitude. For this reason, $\log_{10}$ scaling was applied to all species concentrations, with a floor value of $\varepsilon = 10^{-12}$ to avoid singularities. This transformation brings all variables to comparable orders of magnitude and possesses favorable mathematical properties for PCA applied to compositional data, as recommended in [8]. Following the log-scaling, the variables are standardized by mean-centering and scaling to unit standard deviation. Mean-centering ensures that the principal components capture directions of maximum variance relative to the mean species distribution. In classical PCA reshaped data is scaled column wise. This leads to solving the eigenvalue and eigenvector problem on the correlation matrix. For tensorial data, such a scaling would be dependent upon unfolding direction. Data is therefore scaled to standard deviation globally, that is computed across all spatial locations and temporal snapshots for each species. For combustion data, this is particularly important because it prevents major species with large concentration ranges (H_2, O_2, H_2O) from dominating the decomposition due to their higher magnitude even in log domain, while still preserving relative differences in variability that may carry physical significance.

3. Results and discussion

3.1. Equivalence between PCA and HOSVD

We now establish two fundamental relationships connecting HOSVD to classical PCA through the mode-n unfolding operations defined above.

3.1.1. Equivalence of PCA loadings and U unfolded

PCA loadings are the right eigenvectors of the data tensor $\mathcal{X}$ reshaped to be in the form $\mathbf{X}_{\text{flat}} \in \mathbb{R}^{(\prod_{k\neq n} I_k)\times I_n}$, where rows contain all spatial-temporal configurations and columns represent the I_n features (e.g., species or thermodynamic variables). As long as the eigenvalues are distinct, $\mathbf{X}_{\text{flat}} = \mathbf{X}_{(n)}^T$ and $\mathbf{\Sigma}^T$ contains the same singular values in both cases, the SVD components must satisfy:

$$\mathbf{X}_{(n)}^T = \mathbf{V}_{\text{unfolded}}\mathbf{\Sigma}_{\text{unfolded}}^T\mathbf{U}_{\text{unfolded}}^T = \mathbf{X}_{\text{flat}} \quad (6)$$

Since $\mathbf{X}_{\text{flat}} = \mathbf{U}_{\text{flat}}\mathbf{\Sigma}_{\text{flat}}\mathbf{V}_{\text{flat}}^T$, we have:

$$\mathbf{V}_{\text{unfolded}}\mathbf{\Sigma}_{\text{unfolded}}^T\mathbf{U}_{\text{unfolded}}^T = \mathbf{U}_{\text{flat}}\mathbf{\Sigma}_{\text{flat}}\mathbf{V}_{\text{flat}}^T \quad (7)$$

By uniqueness of the SVD (up to sign), this implies:

$$\mathbf{\Sigma}_{\text{unfolded}} = \mathbf{\Sigma}_{\text{flat}}, \quad \mathbf{V}_{\text{unfolded}} = \mathbf{U}_{\text{flat}}, \quad \mathbf{U}_{\text{unfolded}} = \mathbf{V}_{\text{flat}} \quad (8)$$

Both $\mathbf{\Sigma}_{\text{unfolded}}$ and $\mathbf{V}_{\text{unfolded}}$ are discarded during computation of HOSVD. Therefore the truly relevant result from this is that PCA on $\mathbf{X}_{flat}$, previously done by [1], yields loadings $\mathbf{V}_{\text{flat}}$ that are the left singular vectors $\mathbf{U}_{\text{unfolded}}$ of the standard HOSVD unfolding along the unfolding mode.

3.1.2. Core tensor Frobenius norms equal singular values

For any mode n, the Frobenius norm of each mode-n slice of $\mathcal{G}$ equals the corresponding singular value:

$$\|\mathcal{G}_{i_n=i}\|_F = \sigma_i^{(n)} \quad (9)$$

Unfolding equation (1) along mode n yields:

$$\mathbf{X}_{(n)} = \mathbf{U}^{(n)}\mathbf{G}_{(n)}\mathbf{Q}^{(n)T} \quad (10)$$

where $\mathbf{Q}^{(n)} = (\mathbf{U}^{(n+1)} \otimes \cdots \otimes \mathbf{U}^{(5)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)})^T$ is orthogonal since Kronecker product preserve orthogonality.

Comparing with equation (2), we have:

$$\mathbf{G}_{(n)}\mathbf{Q}^{(n)T} = \mathbf{\Sigma}^{(n)}(\mathbf{V}^{(n)})^T \quad (11)$$

Since both $\mathbf{Q}^{(n)}$ and $\mathbf{V}^{(n)}$ are orthogonal:

$$\mathbf{G}_{(n)} = \mathbf{\Sigma}^{(n)}\mathbf{W}^{(n)} \quad (12)$$

where $\mathbf{W}^{(n)} = (\mathbf{V}^{(n)})^T\mathbf{Q}^{(n)}$ is orthogonal.

The i-th row satisfies $\mathbf{g}_{i,:}^{(n)} = \sigma_i^{(n)}\mathbf{w}_{i,:}^{(n)}$. Since $\|\mathbf{w}_{i,:}^{(n)}\|_F = 1$ (orthogonal rows), we obtain:

$$\|\mathbf{g}_{i,:}^{(n)}\|_F = \sigma_i^{(n)} \quad (13)$$

The i-th row of $\mathbf{G}_{(n)}$ is the vectorization of the mode-n slice $\mathcal{G}_{i_n=i}$, completing the proof.

A more rigorous treatment of these cases can be found in [2], the main takeaway is that HOSVD generalizes PCA to tensors, making it extremely suitable for combustion datasets.

3.2. Equivalence between HOSVD and PCA on data

Figure 1 shows the singular value decay obtained from HOSVD computed with eq. 4. The comparison demonstrates that the singular value decay in both the HOSVD and PCA cases are exaclty the same.

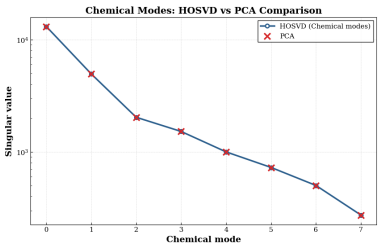

Fig. 1: Generalized singular value decay for chemical dimension of core tensor and singular values from classical PCA

PC loadings and HOSVD modes are as expected completely equal in absolute value. This equivalence, as for classical PCA, is up to signs of eigenvectors/PCs due to basic linear algebra

3.3. Core tensor analysis

Figure 2 shows the decay of singular values along each mode for the system.

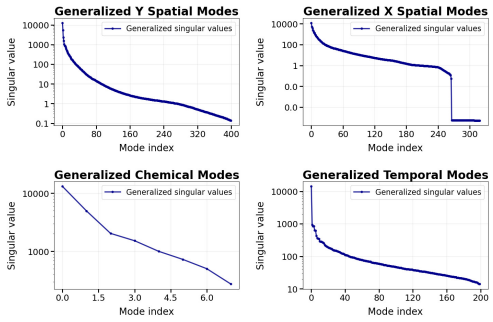

Fig. 2: Generalized singular values decay for HOSVD decomposition. The elbow shape suggests existence of low rank structures in the data

By looking at the elbow in the graphs we can deduce that there is room for significant compression in the data, not only in the dimension identified by PCA but also in the others.

3.4. Compression comparison and Reconstruction error

Without truncation, both PCA and HOSVD are capable of fully recovering the original data, yielding therefore null error. The difference between their performance is therefore to be assessed based on how well they are able to reconstruct the data for a given compression rate, or conversely how much they are able to compress the information within a certain error threshold. The compression ratio, C, is defined as the ratio between the number of elements in the compressed representation and that of the original tensor (or matrix):

$$C = \frac{N_{\text{compressed}}}{N_{\text{original}}}, \tag{14}$$

where $N_{\text{compressed}}$ is the total number of stored elements in the reduced core and mode matrices (for HOSVD) or in the truncated singular decomposition (for PCA). The relative reconstruction error is computed as the normalized Frobenius norm of the difference between the original data tensor $\mathcal{X}$ and its reconstructed approximation $\hat{\mathcal{X}}$:

$$\varepsilon = \frac{\|\mathcal{X} - \hat{\mathcal{X}}\|_F}{\|\mathcal{X}\|_F}. \tag{15}$$

Principal Component Analysis (PCA) truncation can only be applied to the column space of the reshaped data matrix, representing the direction that describes the manifold of chemical evolution within the flames. Despite PCA's proven ability to yield a low reconstruction error, limiting the dimensionality reduction to a single mode (the species direction) constitutes a significant limitation. There is no a priori guarantee that this chosen direction contains the maximum information relevant to the system's low-rank structure. The performance of PCA and HOSVD is typically assessed by iteratively retaining more principal components (PCs) and analyzing the amount of variance (or energy/information) associated with them. Crucially, while PCA truncates only the column space, HOSVD overcomes this limitation by performing truncation along each mode (direction) of the core tensor's singular values, offering a more flexible and comprehensive low-rank approximation of the combustion data. Table 1 reports the compression ratios as defined in 14 and relative reconstruction error, ε

Table 1: Compression ratio and relative reconstruction errors for HOSVD and PCA.

Energy (%)	HOSVD		PCA	
	C	ε (%)	C	ε (%)
83.95	8.05e-6	44.95	0.125	40.06
95.94	1.49e-5	27.68	0.25	20.15
97.96	2.75e-5	21.26	0.375	14.27
99.09	6.26e-5	13.46	0.5	9.52
99.58	1.61e-4	10.73	0.625	6.46
99.98	6.87e-4	6.85	0.75	3.99
99.996	5.15e-3	2.95	0.875	1.91
100	1.0	0.0	1.0	0.0

Due to the application of the chosen energy threshold across all four dimensions of the tensor, rather than just the chemical mode, the resultant reconstruction error for HOSVD is inherently higher than that of PCA for the same threshold. However, the reconstruction error for HOSVD can be arbitrarily improved by iteratively selecting an independent, optimal truncation level for each mode [9].

The most meaningful result evident in Table 1 is that the compression ratio achieved by HOSVD is consistently lower by at least three orders of magnitude for a comparable reconstruction error. This superior efficiency stems from the way each method handles data structure. PCA requires the multi-dimensional data to be reshaped into a matrix, which inextricably mixes the spatiotemporal and chemical information within the same column space. Consequently, when retaining modes for reconstruction, the information contained in the Principal Components (PCs) cannot be reduced optimally. In contrast, HOSVD is applied directly to the original tensor and is able to truncate efficiently along each independent mode (spatial, temporal, and chemical), allowing for a much more compact and efficient low-rank approximation.

4. Conclusion

A robust mathematical connection between widely established and more recent dimensionality reduction techniques is essential for efficient research and to prevent misleading interpretations in complex fields. This work establishes and discusses the relationship between matrix-based PCA and tensor-based HOSVD in the context of combustion data analysis.

We derived a mathematical equivalence between the components of PCA and the corresponding modes of HOSVD, which enables the correct application and future expansion of HOSVD within the combustion community. Numerical validation, performed using a DNS of a hydrogen jet flame, confirmed that HOSVD is significantly more efficient in compressing the multi-dimensional dataset. Specifically, HOSVD achieved comparable reconstruction errors while providing a compression ratio that was lower by at least three orders of magnitude than PCA. These results strongly support the conclusion that tensorial decomposition methods, such as HOSVD, represent a more suitable and efficient approach for analyzing and modeling complex, multi-modal combustion data.

Declaration of competing interest

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

Acknowledgments

IF acknowledges funding from the EU Horizon Europe program (Grant No. 101072779).

References

[1] J. C. Sutherland, A. Parente, Combustion modeling using principal component analysis, Proceedings of the Combustion Institute 32 (1) (2009) 1563–1570.

[2] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM journal on Matrix Analysis and Applications 21 (4) (2000) 1253–1278.

[3] H. Kolla, K. Aditya, J. H. Chen, Higher order tensors for dns data analysis and compression, in: Data Analysis for Direct Numerical Simulations of Turbulent Combustion: From Equation-Based Analysis to Machine Learning, Springer, 2020, pp. 109–134.

[4] W. Austin, G. Ballard, T. G. Kolda, Parallel tensor compression for large-scale scientific data, in: 2016 IEEE international parallel and distributed processing symposium (IPDPS), IEEE, 2016, pp. 912–922.

[5] W. T. Chung, B. Akoush, P. Sharma, A. Tamkin, K. S. Jung, J. H. Chen, J. Guo, D. Brouzet, M. Talei, B. Savard, A. Y. Poludnenko, M. Ihme, Turbulence in focus: Benchmarking scaling behavior of 3D volumetric super-resolution with BLASTNet 2.0 data, Advances in Neural Information Processing Systems (NeurIPS) 36 (2023).

[6] P. Sharma, W. T. Chung, M. Ihme, A physics-informed machine learning approach for predicting dynamic behavior of reacting flows with application to hydrogen jet flames, Combustion and Flame 277 (2024) 114190.

[7] W. T. Chung, M. Ihme, K. S. Jung, J. H. Chen, J. Guo, D. Brouzet, M. Talei, B. Jiang, B. Savard, A. Y. Poludnenko, B. Akoush, P. Sharma, A. Tamkin, Blastnet simulation dataset, uRL: https://blastnet.github.io/ (Jun. 2023).

[8] J. Aitchison, Principal component analysis of compositional data, Biometrika 70 (1) (1983) 57–65.

[9] N. Vannieuwenhoven, R. Vandebril, K. Meerbergen, A new truncation strategy for the higher-order singular value decomposition, SIAM Journal on Scientific Computing 34 (2) (2012) A1027–A1052.