# ON THE USE OF PROJECTION TO LATENT STRUCTURES AND GAUSSIAN PROCESS REGRESSION FOR CHEMISTRY REDUCTION

**Himanshu Dave*, Mohammad Rafi Malik**, Kamila Zdybał*, Hong G. Im**, Alessandro Parente***

himanshu.dave@ulb.be

*Aéro-Thermo-Mécanique (ATM) Department, Ecolé Polytechnique de Bruxelles, Université Libre de Bruxelles (ULB), Brussels, 1050, Belgium

**King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

## Abstract

Combustion chemistry is nonlinear and multiscale. These aspects increase dimensionality of chemical space and computational cost of simulations. In this study, we propose a new, supervised dimensionality reduction technique, called Projection to Latent Structures (PLS), to project chemical species to a reduced space with fewer variables (or scores). In this reduced space, we solve score transport equations, analogous to species transport equations in chemical space. We trained a linear map and a Gaussian Process Regression (GPR) based nonlinear map to predict source terms of the scores. We found that both linear and nonlinear regression maps allow a significant reduction in the number of transported scores. Secondly, we compared the PLS-score and PLS-GPR modeling techniques with previously investigated Principal Component Analysis (PCA) based PC-score and PC-GPR techniques. We observed that PLS-based techniques allow an improved modeling of chemical species than PCA-based techniques, while maintaining comparable training and simulation time. We demonstrated our proposed techniques on methane-oxygen and propane-air datasets from canonical, well-stirred reactor simulations.

## Introduction

Combustion is expected to play a significant and synergistic role in meeting our future energy demands across several sectors [1]. Existing combustion devices thus need to be improved and new combustion modes, like Moderate and Intense Low-oxygen Dilution (MILD) [2, 3], need to be investigated. Numerical simulations play a key role in realizing those improvements and novel technologies [4]. It is well known that combustion chemistry takes a significant fraction of the simulation time. Large computational costs are attributable to aspects of combustion chemistry, such as nonlinear Arrhenius-based source terms in species transport equations, stiffness of governing equations, large spatio-temporal dimensionality, etc. Several species and small time-steps are required to capture complex combustion dynamics. For engineering purposes, the transport equations must be accurately solved on large meshes with time-step size $\left(\Delta t \sim \mathcal{O}(10^{-7} - 10^{-9}\mathrm{s})\right)$. To contribute to the

overarching goal of a better integration of numerical simulations in the design process, accelerating combustion chemistry is, therefore, extremely important.
To accelerate chemistry computations, two broad frameworks are possible – hardware-oriented and modeling-oriented. In [5], the former framework was investigated by taking logarithm of the Arrhenius kinetic models to enable matrix-based formulations and leverage the use of Graphical Processing Units (GPUs). The number of species transport equations is not reduced, rather models are re-formulated to take advantage of better computing architectures. In the latter framework, chemistry computation time is reduced by alternative modeling strategies. Data-driven techniques [6-9] are being increasingly employed for this purpose. Among them, reducing number of transport equations with the help of Principal Component Analysis (PCA) is widely used and established [3, 8, 9].
In this study, we take inspiration from [9] and extend the investigation to the use of Projection to Latent Structures (PLS) [10] for dimensionality reduction. We purport that as PLS focuses on maximizing covariance between chemical variables and their source terms, it extracts a manifold, in a supervised fashion, that is better-informed about the chemical kinetics. Unsupervised PCA extracts a manifold that maximizes explained variance in the chemical space. The species transport equations are then projected onto the manifolds to analogously obtain score transport equations. Given that we require very few scores to be transported, combustion chemistry is reduced in complexity paving way for faster computations.

## Various Machine Learning (ML) techniques

*Principal Component Analysis (PCA)*

PCA is a widely used unsupervised statistical technique [11] that extracts a set of hierarchical principal directions in a dataset by maximizing the variance explained along those directions. The dataset is arranged in a data matrix, $\boldsymbol{X} \in \mathbb{R}^{n_o \times n_c}$, of $n_o$ observations and $n_c$ variables. A covariance matrix, $\boldsymbol{S} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/(n_o - 1)$, is built and eigen-decomposed as $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{L}\boldsymbol{A}^{\mathrm{T}}$, to extract the principal directions. These directions are columns in $\boldsymbol{A}$. Using only the first few ($q$) directions, we perform dimensionality reduction and project $\boldsymbol{X}$ onto another, reduced manifold, $\boldsymbol{A}_q$, and get the Principal Component (PC) scores, $\boldsymbol{Z}_q \in \mathbb{R}^{n_o \times q}$, where $\boldsymbol{Z}_q = \boldsymbol{X}\boldsymbol{A}_q$. Transport equations for $\boldsymbol{Z}_q$ are derived from species transport equations and $\boldsymbol{A}_q$, as shown in Ref. [8].

*Projection to Latent Structures (PLS)*

PLS is a supervised dimensionality reduction technique. It operates on two data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ simultaneously and extracts reduced variables that maximize covariance between them. We performed PLS with chemical species in $\boldsymbol{X}$ and their source terms in $\boldsymbol{Y}$ and used the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [10] to obtain PLS scores $\boldsymbol{T}_q = \boldsymbol{X}_q\boldsymbol{R}_q$. The manifold is stored in $\boldsymbol{R}_q = \boldsymbol{W}_q\left(\boldsymbol{P}_q^{\mathrm{T}}\boldsymbol{W}_q\right)^{-1}$, where $\boldsymbol{W}_q$ and $\boldsymbol{P}_q$ are the weight and loading matrices of $\boldsymbol{X}$. Due to space limitations, we cannot provide complete details here. For details, refer to [10].

*Gaussian Process Regression (GPR)*

GPR does not assume a specific model form and the dependent variables, such as source terms in this study, can be described by a Gaussian distribution with mean function, $m$, and kernel, $K$, as: $\phi \approx GP(m(x), K(x, x'))$.

## Results and Discussion

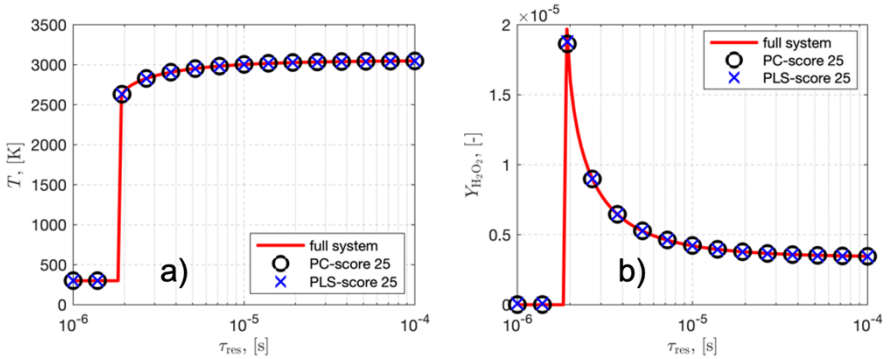


**Figure 1.** a) Temperature, $T$, and b) mass fraction of $H_2O_2$, $Y_{H_2O_2}$, as function of residence time, $\tau_{res}$. PC-score and PLS-score approaches with 25 scores are compared with the detailed solution from Cantera.

| Variable | NRMSE PC-score | NRMSE PLS-score | % Improvement |
|---|---|---|---|
| T | $1.78 \times 10^{-3}$ | $1.73 \times 10^{-3}$ | 2.8 |
| $CH_4$ | $1.26 \times 10^{-3}$ | $9.41 \times 10^{-4}$ | 25.3 |
| OH | $5.32 \times 10^{-3}$ | $3.68 \times 10^{-3}$ | 30.8 |

**Table 1.** Normalized root mean squared error, NRMSE, for different variables with respect to detailed solution. Last column shows improvement in accuracy on using PLS-score instead of PC-score.

Figure 1a shows temperature, $T$, as a function of the residence time, $\tau_{\mathrm{res}}$, of the well-stirred reactor. Profile of $T$ from proposed PLS-score technique (blue crosses) is compared with PC-score (black circles) technique used in [9] and the detailed solution (or full-order solution) from Cantera for $CH_4$-$O_2$ mixture at equivalence ratio, $\phi = 1.0$. Figure 1b shows a similar plot, but for the mass-fraction of $H_2O_2$. In PC-score and PLS-score techniques, the functional relationship between source terms of scores and their corresponding scores ($\boldsymbol{Z}_q$ or $\boldsymbol{T}_q$) is linear. It can be readily inferred that both linear reduced-manifold modeling techniques allow at least 26% reduction (from original 34 chemical species to 25 scores) in chemistry while maintaining proximity to the detailed solutions. While visually, the techniques are comparable to each other, the normalized root mean squared error, NRMSE, between the full-order and reduced-order solutions are lower for PLS-score than PC-score. Last column in Table-1 shows significant improvements of at least 20% over PC-score with the proposed PLS-score technique.

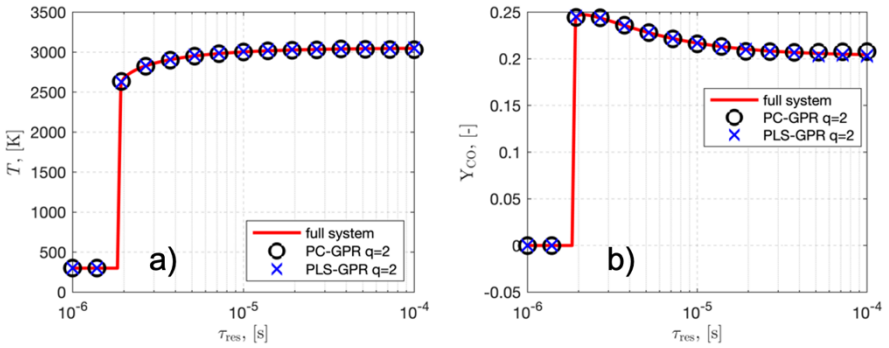


**Figure 2.** Comparison of a) $T$ and b) $Y_{CO}$ profiles with respect to $\tau_{res}$ from proposed PLS-GPR with existing PC-GPR [see Ref. 9] and detailed solution.

| Variable | NRMSE PC-GPR | NRMSE PLS-GPR | % Improvement |
|---|---|---|---|
| T | $6.96 \times 10^{-3}$ | $3.34 \times 10^{-3}$ | 52 |
| $CH_4$ | $2.06 \times 10^{-3}$ | $1.39 \times 10^{-3}$ | 32.5 |
| OH | $1.10 \times 10^{-2}$ | $8.55 \times 10^{-3}$ | 22.3 |

**Table 2.** NRMSE for variables with respect to detailed solution. Last column shows improvement in accuracy with respect to PC-GPR on using PLS-GPR.

Figure 2 and Table 2 show results as shown previously, however, now instead of using linear PCA or linear PLS based relationship between scores and their source terms, we use trained GPR-based regression functions. These regression functions nonlinearly map the scores ($\boldsymbol{Z}_q$ or $\boldsymbol{T}_q$) to their corresponding source terms and closes the set of equations in reduced space. First, it can be readily inferred that with nonlinear, GPR-based functions we see a drastic reduction of number of transported scores from 34 chemical species to just 2 scores. Also, compared to 25 scores with linear approach only 2 scores are transported to obtain NRMSE of same order as PC-score or PLS-score. Secondly, we see a significant improvement in accuracy of 20% and more with respect to PC-GPR as we use PLS-GPR.

To demonstrate the generality of our technique, we used PLS-GPR to reduce combustion chemistry for $C_3H_8$-Air mixture at $\phi = 1.0$. Figure 3 shows the variation of $T$ with respect to $\tau_{\mathrm{res}}$. We used San-Diego reaction mechanism (50 species) in Cantera to model chemistry and obtain detailed solutions. Unlike previous figures, a significant improvement in the accuracy of predicted $T$ profile is visually evident with PLS-GPR. The PLS-GPR predictions (blue crosses) lie closer to the detailed solution (solid red curve) than PC-GPR solution (black circles), and we see a quantitative improvement of nearly 70% over PC-GPR solution.

Thus, we can infer that both supervised PLS and its unsupervised counterpart PCA can reduce number of transport equations to be solved by nearly 90%. However, PLS coupled with nonlinear regression with GPR is shown to significantly improve predictions by at least 20% over PC-GPR technique. The improvements being more

apparent with complex fuels like propane. PLS is therefore a valuable supervised dimensionality reduction technique useful for reducing combustion chemistry. The predicted thermo-chemical variables obtained from score transport equations (on PLS basis) and subsequent reconstruction are closer to the detailed solution.

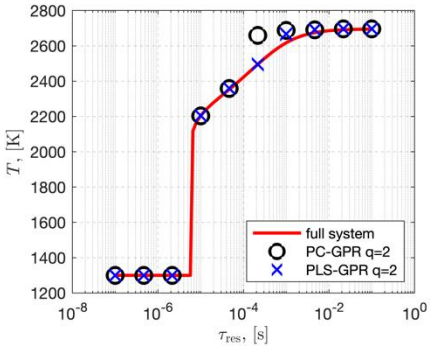


**Figure 3.** Comparison of $T$ with respect to $\tau_{res}$ profiles for existing PC-GPR [see Ref.10], proposed PLS-GPR, and detailed solution from Cantera. The chemical mixture is propane-air at $\phi = 1.0$.

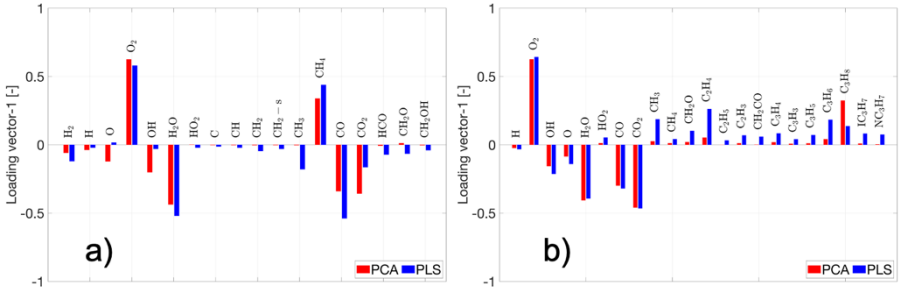


**Figure 4.** Comparison of PCA and PLS loadings on 1st loading vector for a) methane-oxygen and b) propane-air datasets.

To understand possible reasons for observed improvements with PLS, we now investigate the loadings on the direction of reduced manifolds. The first direction (or loading vector) of the manifold identifies dominant structures and consequently most relevant variables. Figure 4 shows bar plots representing the loadings on 1st loading vectors of PCA and PLS-manifolds. PCA identifies species such as fuel ($CH_4$ or $C_3H_8$), $O_2$, CO, $CO_2$, $H_2O$, etc. as important, indicated by greater heights of the red bars. However, PLS vector also have significant loadings on reaction intermediates that appear on fuel break-up, such as $CH_3$, $C_3H_6$, $C_2H_4$, along with species that appear on their eventual oxidation, such as CO and $CH_2O$. This is evident from the heights of the blue bars. It is remarkable that PLS identifies additional species that appear in reaction pathways of methane and propane oxidation without significant intervention from user. We attribute this property of PLS as a reason behind the improvements in prediction accuracy discussed previously.

## Conclusions

In this study, we used a new supervised dimensionality reduction technique, abbreviated as PLS for reducing combustion chemistry. Unlike unsupervised PCA, PLS extracts a manifold that is informed about chemical kinetics. We used methane-oxygen and propane-air mixtures for our investigations. We found that PLS-GPR like PC-GPR can reduce transport equations by nearly 90%. However, PLS-GPR is found to be more accurate (at least 20%) than PC-GPR technique for assessed thermo-chemical variables. The accuracy being more pronounced for propane-air mixture encourages utility for complex fuels. Thus, the newly proposed PLS-GPR, a reduced manifold modeling technique, is favorable for simplifying combustion chemistry and accelerating it. This is attributed to its ability to identify important species and radicals that appear in reaction pathways. The training time for PLS-GPR and PC-GPR were found to be comparable.

## References

[1] Swaminathan, N., "Physical insights on MILD combustion from DNS", *Front. Mech. Eng. 5*, 59 (2019)

[2] Cavaliere, A., & De Joannon, M., "MILD Combustion", *Prog. Energy Combust. Sci. 30*(4) 329-366 (2004)

[3] Parente, A., Sutherland, J. C., Dally, B. B., Tognotti, L., & Smith, P. J., "Investigation of the MILD combustion regime via Principal Component Analysis", *Proc. Combust. Inst. 33*(2) 3333-3341 (2011)

[4] Bilger, R. W., "Future progress in turbulent combustion research", *Prog. Energy Combust. Sci. 26*(4-6) 367-380 (2000)

[5] Barwey, S., & Raman, V., "A neural network-inspired matrix formulation of chemical kinetics for acceleration on GPUs", *Energies 14*(9) 2710 (2021)

[6] Owoyele, O., Kundu, P., Ameen, M. M., Echekki, T., & Som, S., "Application of deep artificial neural networks to multi-dimensional flamelet libraries and spray flames", *International Journal of Engine Research 21*(1) 151-168 (2020)

[7] Parente, A., Sutherland, J. C., Tognotti, L., & Smith, P. J., "Identification of low-dimensional manifolds in turbulent flames", *Proc. Comb. Inst. 32*(1) 1579-1586 (2009)

[8] Sutherland, J. C., & Parente, A., "Combustion modeling using Principal Component Analysis", *Proc. Comb. Inst. 32*(1) 1563-1570 (2009)

[9] Malik, M. R., Isaac, B. J., Coussement, A., Smith, P. J., & Parente, A., "Principal Component Analysis coupled with nonlinear regression for chemistry reduction", *Combust. Flame 187* 30-41 (2018)

[10] Wold, S., Johansson, E., & Cocchi, M, "PLS: partial least squares projections to latent structures", In *3D QSAR in Drug Design: Theory, Methods and Applications* (pp. 523-550) Kluwer ESCOM Science Publisher (1993)

[11] Jolliffe, I. T., & Cadima, J., "Principal Component Analysis: a review and recent developments", *Philos. Trans. Royal Soc. A 374*(2065) 20150202 (2016)