

RESEARCH ARTICLE

A Technique for Characterizing Feature Size and Quality of Manifolds

Elizabeth Armstrong^a and James C. Sutherland^a

^aUniversity of Utah, 201 Presidents' Cir., SLC, UT 84112

ARTICLE HISTORY

Compiled May 13, 2021

ABSTRACT

Effective dimension reduction is a key factor in facilitating large-scale simulation of high-dimensional dynamical systems. The behavior of low-dimensional surrogate models often relies on accurate reconstruction of quantities that can be nonlinear functions of the original parameters. For instance, in low-dimensional combustion models, source terms representative of complex chemical kinetics must be modeled accurately in the reduced dimensional space in order to yield accurate predictions. Features such as sharp gradients or non-uniqueness in a quantity of interest (QoI) may be introduced through a parameterization and pose difficulties for reconstruction techniques. Many existing manifold quality assessments do not consider these features and limit examination to the original parameters and low-dimensional embedding. We have developed a technique for quantitatively assessing manifold quality through characterizing feature size of QoIs by monitoring the change in variance over an increasing filter width. Through identification of variance at small scales, this technique detects undesirable sharp gradients and non-uniqueness of QoIs. Our technique is not limited to a specific reduction method and can be used to compare or assess manifold parameterizations in arbitrary dimensions. We demonstrate our technique on combustion data from both simulation and experiment.

KEYWORDS

dimension reduction; quality assessment; manifold; parameterization; feature size

1. Introduction

Simulation of large-scale high-dimensional dynamical systems, such as combustion, can become computationally prohibitive due to sizable sets of equations to be solved with varying resolution requirements. Dimension reduction is a key aspect in making simulation of such systems more tractable. Many general methods for producing meaningful low-dimensional embeddings exist, including, but not limited to, principal component analysis (PCA) [1], locally linear embedding (LLE) [2], ISOMAP [3], local tangent space alignment (LTSA) [4], and maximum variance unfolding (MVU) [5]. In the context of combustion, parameterizations may also be derived based on intuition of the physics governing the system. Parameterizations common in combustion are discussed more in the following section.

It is essential that surrogate models built over a low-dimensional representation of the original manifold yield accurate predictions for quantities of interest (QoIs)

during a simulation. A particularly important example in combustion simulations is the modeling of source terms, which influence the path a solution takes over a manifold due to chemical reaction. Therefore, the QoIs may include variables other than the original independent variables. Accurate reconstruction of QoIs can become difficult, however, if the dimension reduction introduces sharp gradients or non-uniqueness of values [6].

Many of the existing manifold quality assessments do not take these features into account and limit their examination to the original independent variables and low-dimensional embedding. Moreover, many methods do not apply across reduction techniques. Local continuity meta-criterion [7], trustworthiness and continuity measures [8], and mean relative rank errors [9] are just a few examples of local approaches to quality assessment that consider pairwise distances within a neighborhood of points. While these methods work well for isometric embeddings, they do not accurately assess parameterizations, such as from LLE and LTSA, that distort the geometric structure of each local neighborhood [10]. Many global assessments look at variance computed over a manifold, such as suggested by Tenenbaum et al. [3] or used as criteria in PCA. However, these global measures can ignore small regions of non-uniqueness that may impede accurate QoI-modeling for simulation.

Methods such as those proposed by Zhang et al. [10] attempt to apply more generally across reduction techniques by incorporating both local and global measures that equally pertain to normalized and isometric parameterizations. More recently, assessments based on angles instead of pairwise distances [11] and local rank correlation with special treatment for normalization [12] have been proposed. However, improvements such as these do not address the parameterization of dependent variable QoIs. Furthermore, the optimal parameterization for the independent variables alone is often not optimal for representing quantities that are nonlinear functions of those variables such as the chemical source terms in combustion. While visual inspection of a manifold can provide qualitative assessment for a parameterization, it becomes impractical in more than three dimensions.

In this paper, we introduce a technique that uses information from all QoIs to provide a quantitative quality assessment of a manifold. By monitoring the change in variance for a QoI over an increasing filter width, we can characterize feature sizes over the manifold, which inform how well a surrogate model should represent that QoI. Variance at small scales introduced through a parameterization can indicate undesirable sharp gradients or unacceptable non-uniqueness. This technique provides a quantitative quality assessment that is not limited to a particular reduction method and may be used to compare or assess parameterizations in arbitrary dimensions.

In the following sections, we provide an overview of typical parameterizations for combustion (§2), introduce details on our manifold quality assessment technique (§3), and present results for assessing the parameterizations of interest on combustion data from both simulation and experiment (§4).

2. Parameterizations in Combustion

Manifolds in combustion involving N_s species reacting in a single-phase are uniquely described by a thermochemical state ϕ with $N_s + 1$ variables. This state typically consists of temperature (T), pressure (p), and $N_s - 1$ species mass fractions (Y_i). As combustion can involve hundreds of chemical species participating in thousands of reactions, the original parameterization of a combustion system can become quite

large. The resulting equations can then become prohibitively expensive to solve at large scale as the chemical reactions occur at very small timescales relative to the surrounding flow fields.

Physics-based reduction methods were first used to parameterize combustion manifolds by using a reduced set of meaningful coordinates that capture the underlying combustion processes of interest. Example quantities used in physics-based parameterization include the mixture fraction Z , which provides a measure of the stoichiometry or mixture of fuel and oxidizer; the scalar dissipation rate χ , which provides a measure of the strain on a flame or strength of diffusion; and a progress variable defined by a single or multiple species to track the extent of reaction, to name a few. These lead to common models such as the steady (and unsteady) laminar flamelets [13], flamelet-generated manifolds [14], and flame prolongation of intrinsic low-dimensional manifolds [15]. Many more low-dimensional models for combustion exist that employ simplifying assumptions of the physics which are not demonstrated in this work. A more complete list of manifolds in combustion, physically-derived or otherwise, can be found elsewhere [6]. These models are not always appropriate depending on the physics present; but as the combustion physics becomes more complex, it can become less clear which coordinates should be used to represent the system in the fewest dimensions with the most accuracy. This is especially true given the number of reduction methods that exist for combustion.

Data-based reduction methods have become increasingly popular for deriving low-dimensional representations of combustion data as they remove some of the physical intuition needed for physics-based models. Some examples of data-based methods that have been used to represent combustion systems of varying complexity in a low-dimensional space include ISOMAP [16], proper orthogonal decomposition [17, 18], and many variants of PCA such as global [19–25], local [26–28], and nonlinear [29, 30].

Given multiple approaches to parameterizing a combustion manifold, it can be difficult to assess and compare the quality of all possible parameterizations due to some of the reasons mentioned in §1 as well as the difference in derivation for physics-based and data-based parameters. As an example, we will consider data composed of solutions to the steady flamelet equations [13] for steadily increasing χ until the flame starts to extinguish at which point we incorporate the solution over time to the unsteady flamelet equations until the flame fully extinguishes. The system we consider involves fuel with composition (mole %) 45% CO, 5% H₂, and 50% N₂ in air at 300K and atmospheric pressure. Figure 1a shows the manifold parameterized by (Z, χ) as in steady laminar flamelet modeling. This parameterization is not adequate for describing the data as the transient extinction event that happens once χ becomes sufficiently large contains variation in QoIs that is represented with the single large χ . This can be seen by the vertical plane in Figure 1a between the brown dotted and solid lines highlighting the extinction event at $\ln(\chi) = 6.24$, which is no longer uniquely represented in the two-dimensional $(Z, \ln(\chi))$ plane.

If we instead use a progress variable approach as an improvement for the non-uniqueness seen with (Z, χ) parameters, we achieve the manifold shown in Figure 1b. For this manifold, the progress variable is defined with the species CO₂ as

$$\zeta_{\text{CO}_2} = \frac{Y_{\text{CO}_2} - \min(Y_{\text{CO}_2})}{\max(Y_{\text{CO}_2}) - \min(Y_{\text{CO}_2})}. \quad (1)$$

When we use ζ_{CO_2} instead of χ to parameterize the manifold, the states occurring over the transient extinction event are spread out and uniquely represented on a two-

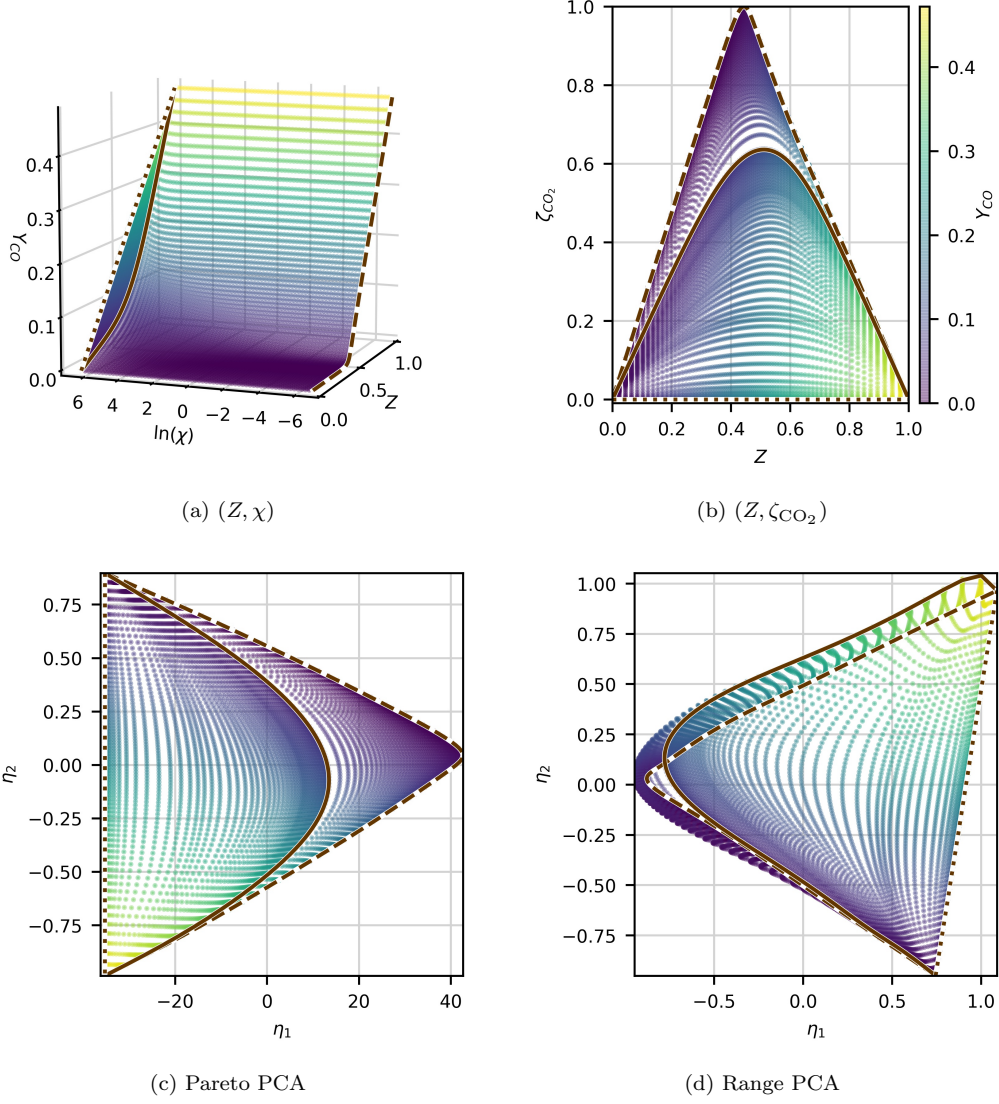


Figure 1. Parameterizations of simulated steady and unsteady flamelet data for CO/H₂ in air colored by Y_{CO} . The brown dashed line indicates the steady solution at the smallest χ of 10^{-3} , the brown solid line displays the steady extinction limit (separation of steady and unsteady states), and the brown dotted line shows the fully extinguished state.

dimensional plane.

Figures 1c and 1d demonstrate two-dimensional parameterizations using PCA, scaling the data by the square-root of the standard deviation (Pareto [31]) and the difference in maximum and minimum values (Range) respectively. It has been shown the resulting structure of the principal components (PCs) defined through PCA is sensitive to how the data is scaled [32]. We use $\vec{\eta}$ to represent the PC parameterization of $\vec{\phi}$, which can then be truncated to an arbitrary dimensionality smaller than $\vec{\phi}$. The PCs are ordered according to how much of the global variance is explained which can guide dimension reduction choices based on the amount of variance to retain over the manifold. However, this is a global criteria, as mentioned in §1, and does not always respect local features. In Figure 1c, we see the Pareto PCA manifold preserves uniqueness of values in two-dimensions; however, based on variance analysis alone, the model

suggests that only one-dimension is needed to retain 99% of the original variance over the manifold. If we truncate the manifold to just one-dimension, we will introduce non-uniqueness in our parameterization. Therefore, the global analysis PCA provides falls short in this instance for determining an adequate dimension for parameterization. In general, the global PCA criteria for choosing a dimensionality of highly nonlinear manifolds, which is based on a linear projection, may not be adequate.

Figure 1c also shows the Pareto PCA reduction performing at least as well as the physics-based (Z, ζ_{CO_2}) reduction. However, this will not always be true for all PCA scalings. Range PCA, shown in Figure 1d, has introduced non-uniqueness in the regions occurring before and after the steady extinction limit through projecting the original manifold into two-dimensional PC space. This would suggest three or more dimensions are needed to adequately represent the manifold with Range scaling compared to Pareto; and this analysis might change depending on the QoI in question as well, which we want to take into account when assessing quality of manifold parameterizations.

We note that if the transient extinction data were not included in Figure 1a, then the QoI data would be unique over the (Z, χ) manifold. In many of the parameterization methods for combustion, surrogate models for QoIs are built using data generated uniquely over the chosen parameters. For those cases, it may not make sense to check for non-uniqueness on the generated manifolds. However, it can still be useful to have a method for comparing feature sizes of parameterizations such as these and assessing their applicability to complex scenarios of interest using representative experimental or simulated data.

The data generated for the demonstration shown here allows us to easily assess manifold quality for various two-dimensional parameterizations through visual inspection. However, it is impractical to visually assess each kind of parameterization for each QoI, and this qualitative analysis is limited to three dimensions while more may be needed in representing more complex data. We note that methods for quantitative comparisons of parameterizations in combustion using a ground truth from direct numerical simulation exist [33, 34], but in general we do not want to rely on the necessity for a “truth” when comparing reduction methods. This motivates the need for a quantitative assessment that allows us to compare reduction methods for all QoIs efficiently. We introduce our proposed solution for manifold quality assessment in the following section and later apply the technique in comparing the parameterizations discussed here on both computational and experimental combustion datasets in varying dimensions.

3. Manifold Quality Assessment

The motivation for the technique proposed here comes from needing to assess various projections of high-dimensional data onto a set of parameters for building surrogate models of QoIs. A manifold maintaining uniqueness and moderate gradients in QoI values should produce more accurate models. We therefore aim to characterize feature size with the focus on identifying smaller scales that present more difficulties for accurate regression. In the following subsections, we present the formulation of the technique (§3.1) and demonstrate how scales of variation are detected (§3.2) and can be related to non-uniqueness in QoIs (§3.3).

3.1. Formulation

In order to determine feature sizes for QoIs over a manifold, we analyze a normalized variance computed over varying filter widths or bandwidths. Through this measure, we can detect scales of variation below a bandwidth and monitor how the variation changes as the bandwidth is increased. This is similar to how Gaussian filtering is used to reduce noise below a filter width in signals or images, but with our focus on the scales being filtered. To ensure that a bandwidth has the same effect in each dimension of the manifold, which may exist over varying magnitudes, the independent variable observations are centered and scaled between zero and one as follows

$$\hat{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2)$$

where x_i , being a vector of independent variable coordinates, represents a point on the original manifold and $\hat{x}_i$ then corresponds to the normalized manifold. This is equivalent to applying an anisotropic filter/kernel on the original manifold.

For a given QoI, y , we then compute a filtered variance along the manifold using a specified bandwidth, σ , and normalize by the global variance. This defines a normalized variance as

$$\mathcal{N}(\sigma) = \frac{\sum_{i=1}^n (y_i - \mathcal{K}(\hat{x}_i; \sigma))^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

where n is the number of observations, $\bar{y}$ is the arithmetic average over the whole manifold, and $\mathcal{K}$ is the weighted average quantity calculated using Nadaraya-Watson kernel regression [35]:

$$\mathcal{K}(\gamma; \sigma) = \frac{\sum_{j=1}^n \mathcal{W}_j(\gamma; \sigma) y_j}{\sum_{j=1}^n \mathcal{W}_j(\gamma; \sigma)} \quad (4)$$

for a Gaussian kernel $\mathcal{W}_j$, defined for each point $\hat{x}_j$ on the normalized manifold, which can be evaluated given a bandwidth σ as

$$\mathcal{W}_j(\gamma; \sigma) = \exp\left(\frac{-\|\hat{x}_j - \gamma\|_2^2}{\sigma^2}\right). \quad (5)$$

We use a Gaussian kernel for its desirable properties such as easily extending to arbitrary dimensions, offering infinite rather than compact support, and applying to point-cloud as well as structured data.

The normalized variance, $\mathcal{N}(\sigma)$, represents variance in a QoI at scales below a bandwidth, σ . For a manifold with unique observations, $\mathcal{N}$ will approach zero as the bandwidth approaches zero and will approach one as the bandwidth increases to encompass the manifold. More specifically,

$$\lim_{\sigma \rightarrow 0} \mathcal{N}(\sigma) = \begin{cases} 0 & \text{if all } \hat{x}_i \text{ are distinct} \\ \frac{\sum_{i=1}^n (y_i - \bar{y}_i^0)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} & \text{if } \exists \hat{x}_i = \hat{x}_j \text{ for } i \neq j \end{cases} \quad (6)$$

where $\bar{y}_i^0$ is the arithmetic average of all observations of a QoI y occurring at the same location as $\hat{x}_i$. Therefore, if all independent coordinates are distinct, then $\bar{y}_i^0 = y_i$ and

$\lim_{\sigma \rightarrow 0} \mathcal{N}(\sigma)$ reduces to zero, as shown in (6). At the other extreme,

$$\lim_{\sigma \rightarrow \infty} \mathcal{N}(\sigma) = 1 \quad (7)$$

since $\mathcal{K}$ will approach the global arithmetic average $\bar{y}$ with σ large enough. The normalized variance will smoothly and monotonically increase between these bounds since the Gaussian kernel is strictly positive and has infinite support. Larger values of σ at which this rise occurs indicate larger scales of variation in the QoI which should better facilitate regression over the manifold.

3.2. Detecting Scales of Variation

We will now demonstrate how the technique proposed above can detect scales of variation in data through an example problem. We start by considering three signals $\forall x \in [0, 1]$ as

$$w_1(x) = \sin(2\pi x) \quad (8)$$

$$w_2(x) = w_1(x) + \sin(4\pi x) \quad (9)$$

$$w_4(x) = w_2(x) + \sin(100\pi x) + \sin(1000\pi x) \quad (10)$$

so that w_1 is composed with one frequency, w_2 contains two superimposed frequencies (one of which matches w_1), and w_4 is made of four superimposed frequencies with two matching those of w_2 . These signals are shown in Figure 2 evaluated at 5000 uniformly spaced points in the domain.

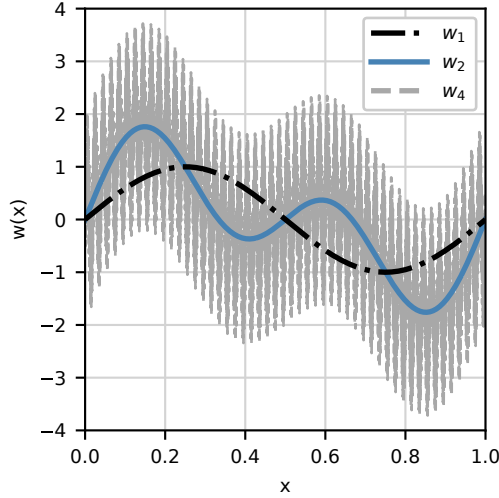


Figure 2. Signals given by (8)-(10) evaluated at 5000 uniformly spaced points in the domain.

We now want to determine if the normalized variance technique can pick up the multiple frequencies, or scales over which the signal is varying, present in the signals of Figure 2. Through steadily increasing bandwidths the normalized variance monitors the increase in the variance until the global scale is reached. The result is shown in Figure 3, where we see the expected behavior of an increase of the normalized variance from zero at small bandwidths to one at a bandwidth corresponding to the

global variance. The normalized variance curves in this figure, therefore, represent the cumulative fractional variance explained by scales less than σ . The curve corresponding to w_1 , which only has one frequency, has a single, smooth rise which starts to occur at a larger scale ($\sigma \approx 0.2$) than the others due to the smaller frequency. On the other hand, the curve for w_4 has four separate humps in the normalized variance, where the two corresponding to the highest frequencies are more distinct than the two for the lower frequencies. This is due to the larger frequencies being further apart than the smaller frequencies which end up contributing to variance over similar scales. We can see this blending in scales of variation occurring for the w_2 signal as well. The scales associated with the highest frequencies in w_4 (showing up for σ between 10^{-4} and 10^{-2}), visually corresponding to the “noise” in the signal of Figure 2, will be more difficult to accurately model than the smaller frequencies. Having the scales of variation span orders of magnitude as seen in w_4 can also make accurate modeling difficult.

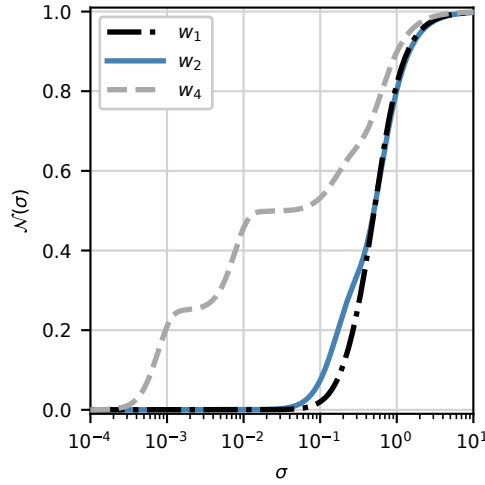


Figure 3. Normalized variance computed using (3) for w_1 , w_2 , and w_4 of (8)-(10).

We also note that the fastest changes in $\mathcal{N}$ of Figure 3 are occurring at σ near the intrinsic frequencies of the signals. For example, the scale pointed out earlier for w_1 around $\sigma \approx 0.2$ is close to $1/(2\pi)$. In order to better highlight these points, we look at the derivative of $\mathcal{N}$ over the logarithmically scaled bandwidths, corresponding to the display of $\mathcal{N}$ in Figure 3, which relays how fast the variance is changing as the bandwidth changes. More specifically, we compute $\hat{\mathcal{D}}(\sigma)$ from

$$\mathcal{D}(\sigma) = \frac{d\mathcal{N}(\sigma)}{d\log_{10}(\sigma)} + \lim_{\sigma \rightarrow 0} \mathcal{N}(\sigma) \quad (11)$$

and

$$\hat{\mathcal{D}}(\sigma) = \frac{\mathcal{D}(\sigma)}{\max(\mathcal{D}(\sigma))}. \quad (12)$$

In (11), we add $\lim_{\sigma \rightarrow 0} \mathcal{N}(\sigma)$ to the derivative in order to account for any overlapping points that may cause $\mathcal{N}$ to level off at a nonzero value for small bandwidths. We also

normalize to better compare highlighted scales among the signals, or in general among manifold parameterizations. In practice, we approximate the derivative in the first term of (11) with central finite differencing and the second term with $\mathcal{N}(\sigma = 10^{-16})$. This results in smooth curves for sufficiently sampled bandwidths.

The results for computing $\hat{\mathcal{D}}(\sigma)$ using (12) for w_1 , w_2 , and w_4 are presented in Figure 4. This separates scales of variation even more clearly through local maxima observed in $\hat{\mathcal{D}}(\sigma)$. Even for the similar frequencies of w_2 , separate peaks can now be observed, and the amount of blending that occurs over certain scales of variation is conveyed through the blending of peaks in the derivative. We can also see how the rate of change in $\mathcal{N}$ starts to pick up around the intrinsic frequencies at $\sigma \approx 1/(1000\pi), 1/(100\pi), 1/(4\pi)$, and $1/(2\pi)$, and peaks around the same order of magnitude. This demonstrates how our normalized variance can detect intrinsic scales of variation in a manifold.

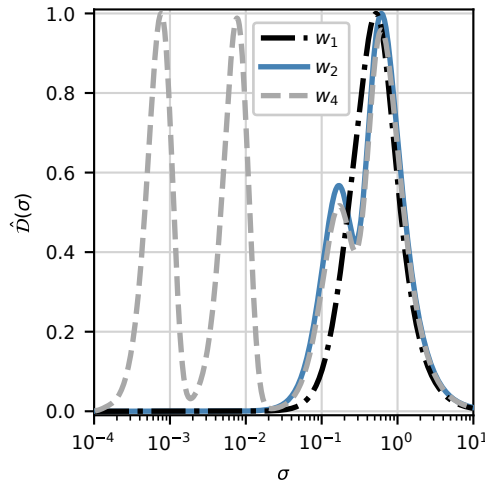


Figure 4. $\hat{\mathcal{D}}(\sigma)$ computed using (12) for Figure 3.

Analysis of the behavior of $\hat{\mathcal{D}}(\sigma)$ suggests an approximation of a Fourier transform. For a single scale of variation, as found in w_1 , we get a single peak which behaves like a smoothed delta function from a Fourier transform. For multiple scales of variation, multiple peaks then materialize. However, unlike a Fourier transform, our technique easily extends to point-cloud data and arbitrary dimensions. We now want to apply the analysis in the context of dimension reduction and determine whether scales of variation are characteristic features of a manifold or artifacts of the reduction. The following section will demonstrate this extended analysis for detecting scales of non-uniqueness in QoIs.

3.3. Quantitative QoI-Dependent Quality Assessments

In this section we will demonstrate how our technique is QoI-dependent and can distinguish between characteristic features and non-uniqueness or overlapping states on a manifold. We first create a two-dimensional helical manifold defined by independent

variables u and v as

$$\left. \begin{aligned} u(t) &= (1-t) \cos(48\pi t) \\ v(t) &= (1-t) \sin(48\pi t) \end{aligned} \right\} \forall t \in [0, 1] \quad (13)$$

then define a dependent variable f for a specified parameter m as

$$f(u, v; m) = \sin(2\pi(u^2 + mv)). \quad (14)$$

By changing the parameter m , we can create multiple dependent variables or QoIs that have various degrees of contribution from v . The top row of Figure 5 shows the resulting manifold, created using 5000 uniformly spaced samples of t , colored by three separate QoIs for $m = 0, 0.1$, and 1 . The manifold is two-dimensional and cannot be projected onto either u or v while maintaining uniqueness in the independent coordinates. However, we are interested in the variation of QoIs over the manifold, so non-uniqueness is QoI-dependent. This is demonstrated in the bottom row of Figure 5 which presents the scatter in QoIs for a particular value of m when the manifold is projected onto u . In the case for $m = 0$, f has no dependence on v and can be projected onto u while maintaining uniqueness. As we start to increase the value of m , moving to the right in Figure 5, more and more non-uniqueness in f shows up due to the projection into one-dimension. Therefore, while a one-dimensional parameterization for $f(u, v; 0)$ on u is acceptable, it is not for $m = 0.1, 1$.

Results for applying our quality assessment technique to each QoI over the one-dimensional manifold defined by u and original two-dimensional manifold are shown in Figure 6. Figure 6a displays the normalized variance and Figure 6b presents the corresponding $\mathcal{D}(\sigma)$ computed using (12). For the QoI with $m = 0$, we see a single rise in the normalized variance of Figure 6a and the corresponding single peak in Figure 6b for both the 1D and 2D manifolds. This identifies a single feature size for $f(u, v; 0)$, which corresponds to maintaining uniqueness in the projection onto u . This feature size is around $\sigma \approx 0.1$, which can be rescaled to 0.2 on the original manifold (which varies from -1 to 1), highlighting the dominant length scale of variation in $f(u, v; 0)$.

As m is increased to 0.1 , overlapping states are introduced in the one-dimensional projection onto u . We can see two scales of variation being detected in Figure 6 around order 10^{-4} and 10^{-1} , with more emphasis given to the larger scale. In comparison, only one scale of variation is found for $m = 0.1$ over the two-dimensional manifold. Therefore, a smaller scale of variation is being introduced through the projection of $f(u, v; 0.1)$ onto u ; and we know this additional scale of variation can be attributed to the non-uniqueness being injected through nonzero m . As we increase m to 1 , the degree of non-uniqueness increases further and more emphasis is given to the smaller scale associated with overlap on the 1D manifold.

We also note that the normalized variance for these 1D projections of nonzero m approaches zero for small bandwidth despite there being non-uniqueness demonstrated in the scatter plots of Figure 5. We know from (6) that for non-unique observations, the normalized variance should approach a nonzero value as the bandwidth decreases toward zero; however, this may be difficult to capture with discrete data points as used in this example. Therefore, the location of the peaks around $\sigma \approx 10^{-4}$ in Figure 6b are a byproduct of spacing in t rather than a characteristic feature size. If the resolution in t is changed (while maintaining enough resolution to capture the characteristic feature size around $\sigma \approx 10^{-1}$), the location of the second peak introduced by the one-dimensional projection will change while the peak at the larger σ will remain constant.

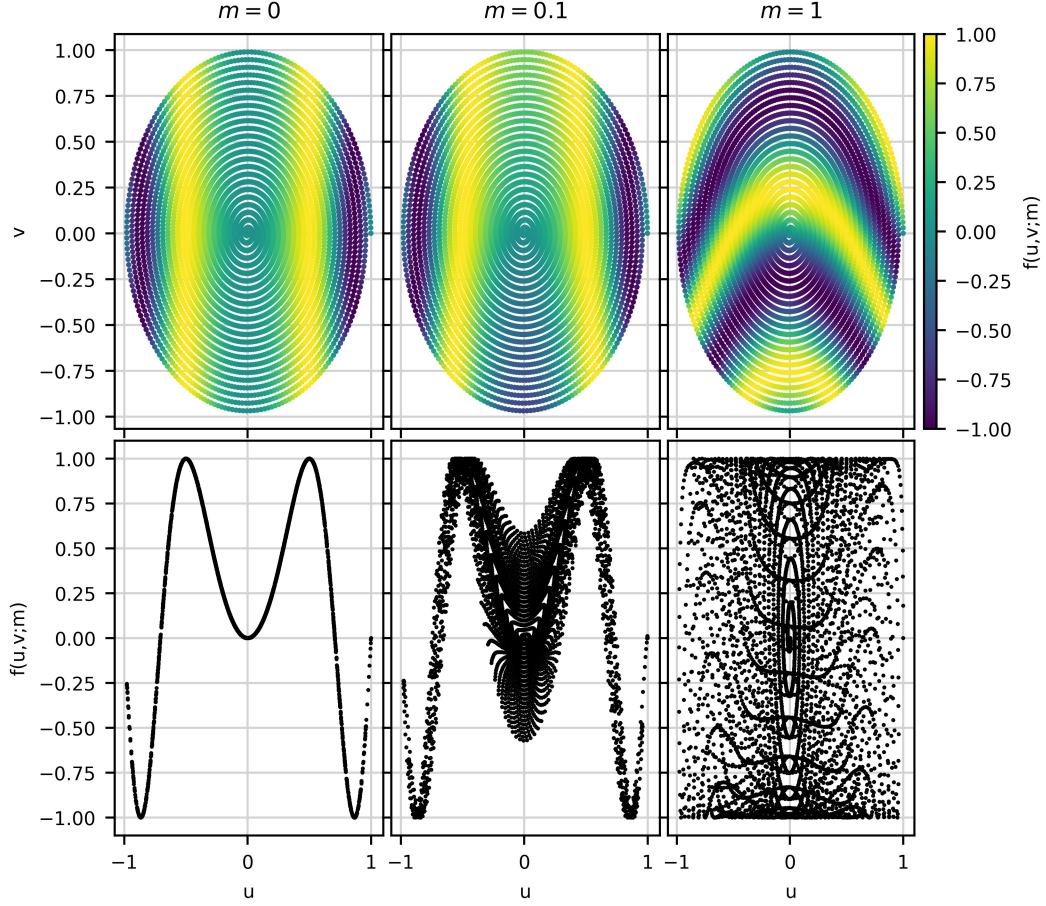


Figure 5. QoIs from (14) over the manifold defined in (13) evaluated at 5000 uniformly spaced values of t . The top row shows the QoIs over the two-dimensional manifold and the bottom row shows the QoIs projected onto u alone.

This is an attribute we can use to distinguish between variation from characteristic features and variation from non-uniqueness.

Since, in some cases, it may be difficult to increase the resolution of datasets, we will explore the sensitivity of peak locations in $\hat{\mathcal{D}}$ to data spacing through random subsampling of the original data. The peak locations should be more sensitive to the data samples for peaks corresponding to overlap than for peaks corresponding to characteristic features. Figure 7 demonstrates this sensitivity for $f(u, v; 0.1)$ and $f(u, v; 1)$ projected onto u . Each curve in Figure 7 is an average of five realizations of $\hat{\mathcal{D}}$ for a randomly selected sample based on a percentage of the original dataset. We can see as less of the original data is used to compute $\hat{\mathcal{D}}$, the peaks corresponding to overlap shift further to the right (due to larger spacing between points) while the peak for the characteristic feature remains constant. The peaks around $\sigma \approx 10^{-4}$ in Figures 7a and 7b experience an 18% and 15% shift to the right, respectively, in $\log_{10}(\sigma)$ from the full dataset to a randomly selected 25% of the dataset.

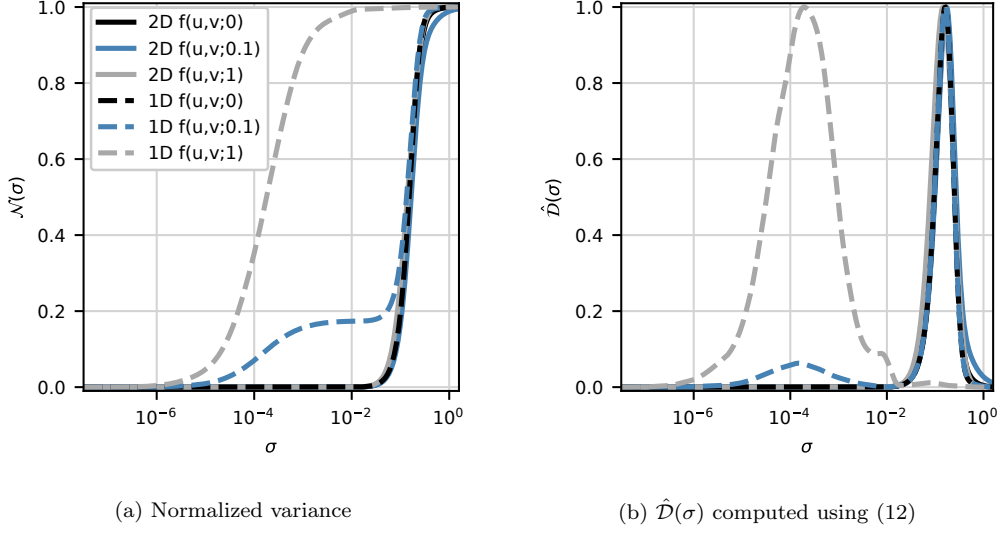


Figure 6. Quality assessment for $f(u, v; m)$ in (14) over the original two-dimensional manifold and one-dimensional manifold projected onto u .

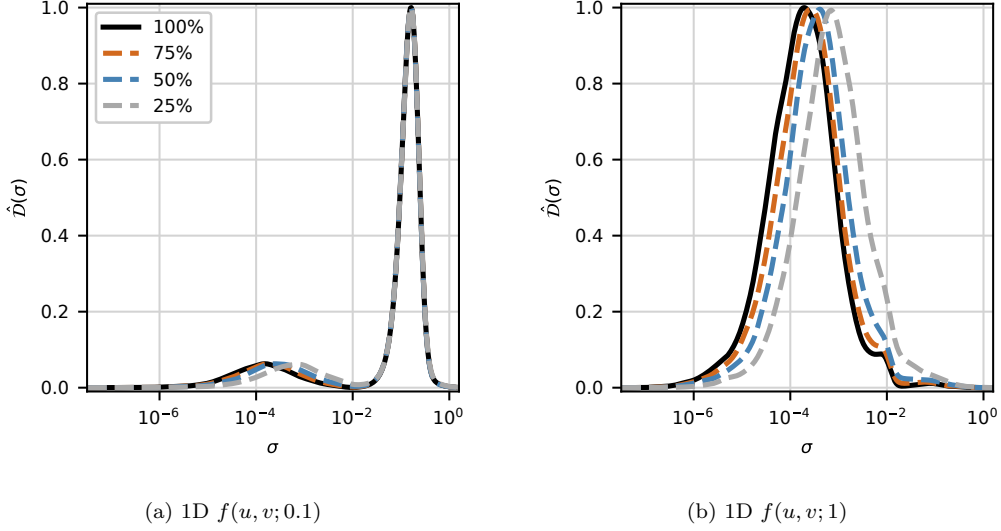


Figure 7. Curves for $\hat{D}$ averaged over five realizations of a randomly selected sample based on a percentage (specified in the legend) of the original one-dimensional dataset. The 100% label corresponds to the full one-dimensional datasets shown in Figure 6b.

In order to ensure this method of subsampling can be used to distinguish between small scales of non-uniqueness and small features, we apply this technique to the w_4 QoI from (10) in §3.2. The w_4 signal in Figure 4 also showed variation at small scales for a one-dimensional manifold, but the values over the single independent variable were uniquely represented. Figure 8 shows the resulting peaks in $\hat{D}$ for subsampling the w_4 dataset. We can see most of the peak locations remaining the same, while the peak with the smallest σ experiences a slight shift as the percentage of data used is decreased. This small shift is due to under-resolving the variation occurring at the largest frequency, but is much less sensitive to data spacing than in the presence of overlap. This is seen with the peak only shifting by $\approx 1\%$ in $\log_{10}(\sigma)$ compared to

the 15-18% shift seen in Figures 7a and 7b due to non-uniqueness. We can therefore use this method of estimating sensitivity of scales of variation to data spacing through subsampling in order to distinguish between non-uniqueness and characteristic features of a manifold.

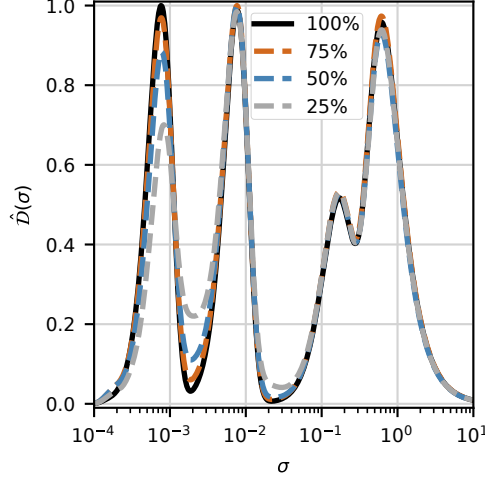


Figure 8. $\hat{D}$ averaged over five realizations of a randomly selected sample based on a percentage (specified in the legend) of the original w_4 dataset from Figure 2.

This example demonstrates how our technique provides a quantitative assessment of parameterizations for QoIs through feature size. The σ locations for peaks in $\hat{D}$ provide a measure of scales at which each QoI is varying fastest (in a normalized space), which can be related to feature sizes. Parameterizations that give rise to larger feature sizes are preferred as those features should be easier to accurately model. Scales of variation due to non-uniqueness in QoIs over a manifold are also detected, and can be distinguished from characteristic features through their high sensitivity to data spacing. Capturing percent changes in peak locations due to subsampling a dataset provides an estimate of this sensitivity. Finally, this example showed how both manifold coordinates and QoIs are incorporated in defining non-uniqueness and determining the validity of parameterizations.

4. Application to Combustion

We will now use the quantitative manifold quality assessment technique outlined in §3 to compare parameterizations of combustion data for applicable QoIs. The technique will be demonstrated on both simulated (§4.1) and experimental (§4.2) datasets.

4.1. Simulated Data

We will start with comparing the four two-dimensional parameterizations for describing the simulated steady and unsteady CO/H₂ flamelet data displayed in Figure 1. This provides an opportunity to validate our quality assessment technique with the visual assessment performed in §2 before moving to a more complicated dataset. Our representative QoIs for analysis will include temperature, select species (Y_{CO} , Y_{H_2} ,

Y_{CO_2} , and Y_{OH}), and applicable source terms. For single phase systems, such as we are dealing with, equations for the mixture fraction do not contain a source term. Therefore, for the (Z, χ) parameterization, since χ is typically computed from Z rather than transported in laminar cases, no source terms will be considered in the manifold assessment. For the (Z, ζ_{CO_2}) representation, a source term, S_{CO_2} , for the transport of Y_{CO_2} will be considered. When using PCA parameterizations, the same transformation for $\vec{\phi}$ to $\vec{\eta}$ applies to the source terms of $\vec{\phi}$. Therefore, both the Pareto and Range PCA parameterizations will include source terms S_{η_i} for each η_i in the assessment.

Figure 9 shows the computed $\hat{D}$ for all the QoIs on each two-dimensional manifold from Figure 1. We see that the curves in Figure 9a for the (Z, χ) parameterization never actually reach zero for all QoIs. This arises from $\lim_{\sigma \rightarrow 0} \mathcal{N}(\sigma)$ not reaching zero due to variation in the QoIs during the extinction event being projected onto the same set of grid points in Z for a single value of χ . Therefore, (Z, χ) is an inadequate parameterization for the combined steady and unsteady flamelet data.

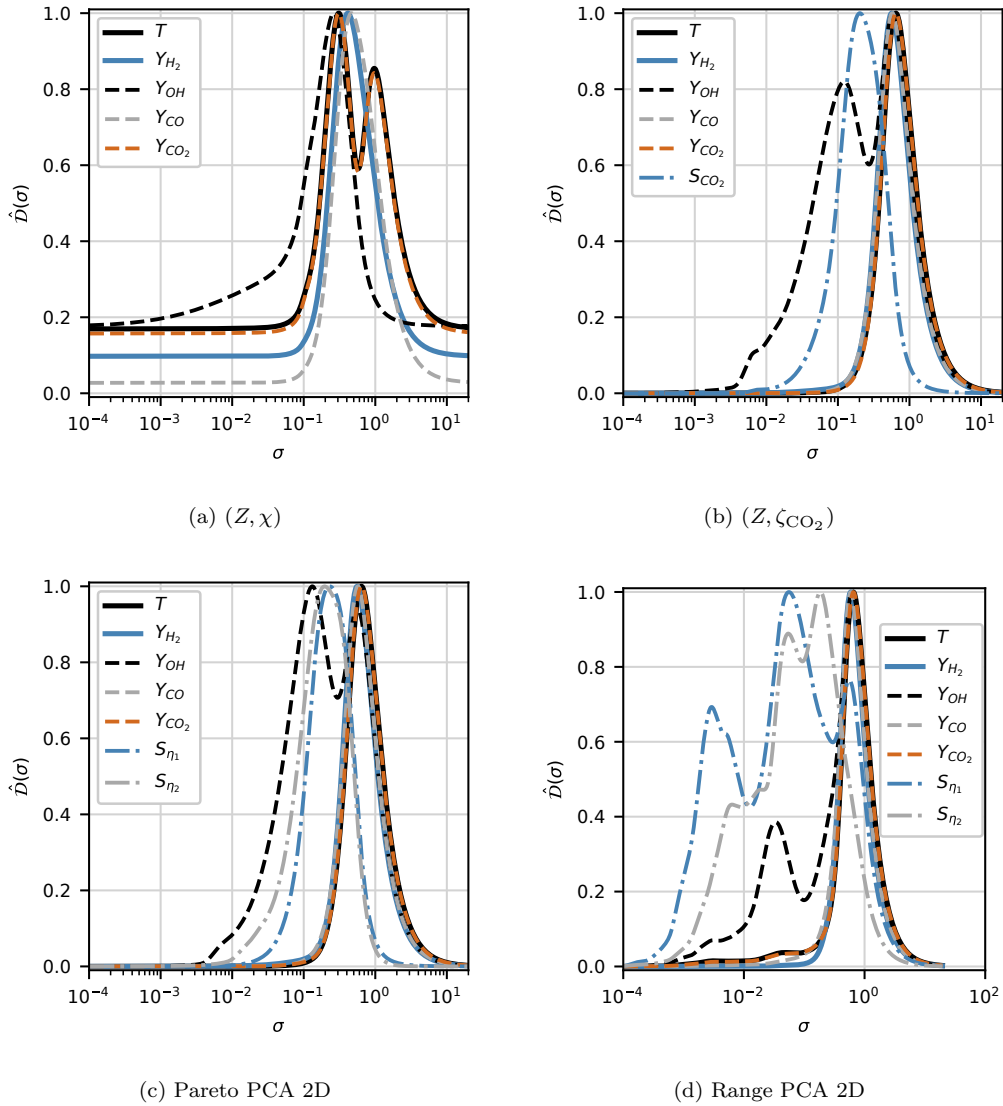


Figure 9. $\hat{D}(\sigma)$ for relevant QoIs on the manifolds of Figure 1.

In Figures 9b - 9d, we can see that out of all the QoIs, source terms and the radical species Y_{OH} show the smallest scales of variation, and in many cases show more than one. In general, source terms tend to be difficult to parameterize well as they are nonlinear functions of the original coordinates in high-dimensional space and radical species can be difficult to model as they tend to evolve over smaller time and length scales during the course of combustion than major species. We also note that Pareto PCA and the progress variable parameterization show the QoIs with larger, more unique feature sizes than Range PCA. This is due to the Range PCA manifold folding over itself in two-dimensions which leads to the smaller scales of variation and large range over which variation is observed for a given QoI. Despite the two-dimensional Range PCA manifold introducing overlapping states, its curve in Figure 9d reaches zero at small σ because the observations do not lie directly on top of each other as they do on the (Z, χ) manifold. Therefore, the overlap that can be seen in Figure 1d has to be detected through further analysis.

Since our technique applies to arbitrary dimensions, we can also compare the scales of variation found in Figure 9 for two-dimensional projections of the simulated flamelet data with the scales of variation over the original 12-dimensional space in Figure 10. This figure shows that all QoIs exhibit one dominant feature size in the original 12-dimensional space and that this feature size is larger than on the two-dimensional manifold projections. Therefore, the projection into two dimensions is leading to sharper gradients in QoIs compared to the original full-dimensional space and in some cases is introducing new scales of variation, such as for Y_{OH} . While sharpening gradients is undesirable, it is still possible to build accurate models for those smaller features unlike when non-uniqueness is introduced by the projection. We therefore need to determine whether these new scales are features that need to be accurately modeled or represent variation from non-uniqueness in which case the parameterization is inadequate for modeling the QoIs.

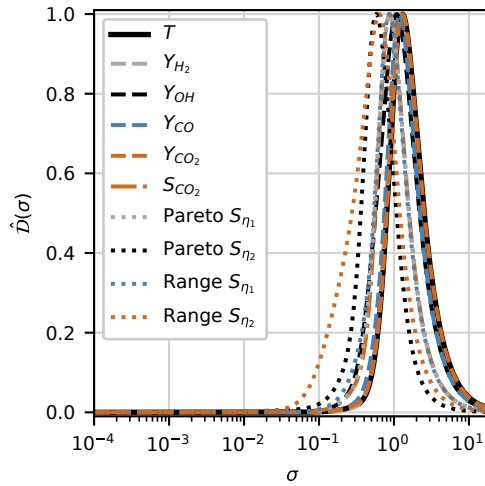


Figure 10. $\hat{D}(\sigma)$ for all QoIs in the original 12-dimensional space for the simulated flamelet data.

We will use sensitivity of $\hat{D}$ peak locations (σ_{peak}) to data subsampling as outlined in §3.3 in order to differentiate between non-uniqueness and characteristic features on the two-dimensional manifold projections. Figure 11 displays the effect of random subsampling on $\hat{D}$ for Y_{OH} over the (Z, ζ_{CO_2}) and Pareto PCA 2D manifolds. We find

that subsampling down to 25% of the two-dimensional datasets does not significantly impact the location of the peaks for either parameterization. There is only an $\approx 1\%$ shift in $\log_{10}(\sigma_{\text{peak}})$ for both manifolds. This suggests that the secondary peak introduced in the two-dimensional projections is simply an additional gradient that can be modeled rather than non-uniqueness created by a fold in the manifold. We can confirm this visually in Figures 1b and 1c as well.

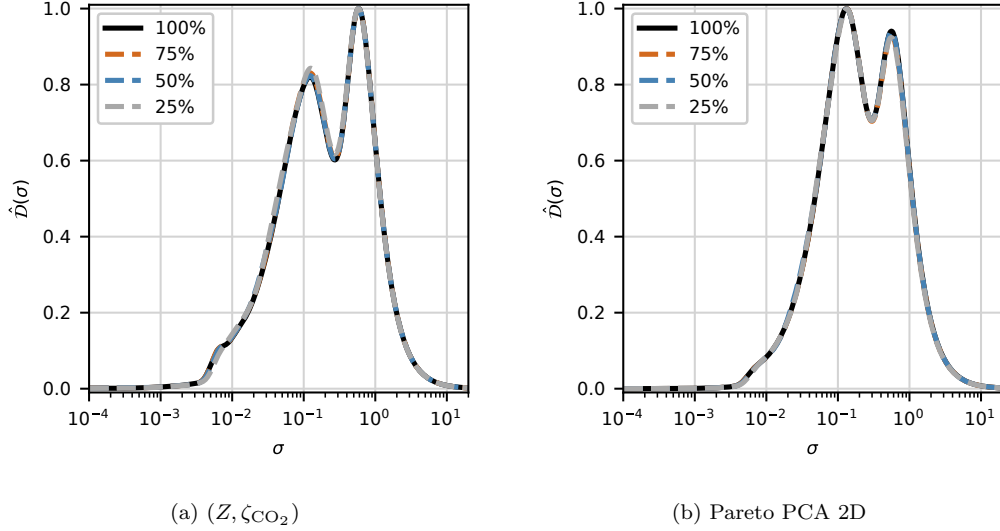


Figure 11. Curves for $\hat{\mathcal{D}}$ of Y_{OH} averaged over five realizations of a randomly selected sample based on a percentage of the two-dimensional dataset.

In contrast, we find that the additional peaks in $\hat{\mathcal{D}}$ introduced through the Range PCA 2D projection show high sensitivity to data subsampling. Figure 12 displays this sensitivity of $\hat{\mathcal{D}}$ to random sampling for QoIs on the Range PCA 2D manifold that had secondary peaks introduced compared to the original high-dimensional space in Figure 10. The two largest peaks in each subfigure of Figure 12, actually show similar behavior seen in Figure 11 with low sensitivity to data spacing. However, the difference between the Range PCA 2D manifold and the Pareto PCA 2D and (Z, ζ_{CO_2}) manifolds is that more peaks have been introduced at even smaller scales when using Range PCA. These additional peaks show high sensitivity to subsampling, with $\approx 17\%$ shifts in $\log_{10}(\sigma_{\text{peak}})$ or in some cases, like Figure 12c, a combination of two smaller peaks into one. This high sensitivity to data spacing indicates non-uniqueness in the projection, which we can visually confirm in Figure 1d. Since the source terms on the Range PCA 2D manifold experience the most variation right around the fold in the manifold, it also makes sense that they would show the most sensitivity to data sampling. Therefore, our manifold assessment technique is able to correctly determine the Range PCA 2D parameterization as inadequate for describing the simulated flamelet data.

We can now further analyze the two acceptable two-dimensional parameterizations of Pareto PCA and (Z, ζ_{CO_2}) by comparing the peak locations for all QoIs. Figure 13 shows the peak locations on a logarithmic scale for each QoI on the manifolds. We note that two bars exist for Y_{OH} due to the existence of two peaks as shown in Figure 11. This essentially provides a comparison of feature sizes on the normalized two-dimensional manifold projections. We can see that there isn't much of a difference in these QoI feature sizes for the Pareto PCA and (Z, ζ_{CO_2}) manifolds. Further comparison, if desired, can consider other case-specific factors. For example, having

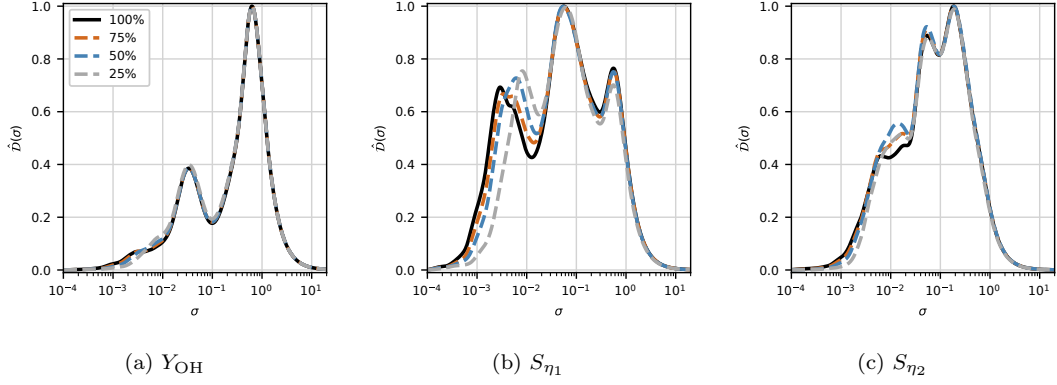


Figure 12. Curves for $\hat{D}$ on the Range PCA 2D manifold for various QoIs averaged over five realizations of a randomly selected sample based on a percentage of the two-dimensional dataset.

fewer source terms to model might be considered favorable in some cases. Our manifold quality assessment has shown that both of the two-dimensional Pareto PCA and (Z, ζ_{CO_2}) parameterizations are acceptable and result in similar QoI feature sizes.

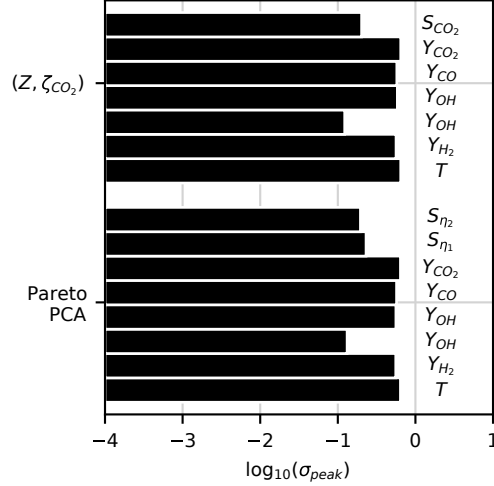


Figure 13. Feature sizes associated with the (Z, ζ_{CO_2}) and Pareto PCA parameterizations from Figure 1 using the proposed quantitative manifold assessment. The bar represents the locations in σ where peaks in $\hat{D}$ occur for each QoI. From bottom to top in each parameterization set, the bars correspond to T , Y_{H_2} , Y_{OH} , Y_{CO} , and Y_{CO_2} followed by relevant source terms. Note that two bars exist for Y_{OH} corresponding to the two peaks present in $\hat{D}$ shown in Figure 11.

With our manifold quality assessment technique, we were able to determine that out of the four parameterizations of simulated flamelet data shown in Figure 1, only the Pareto PCA and (Z, ζ_{CO_2}) parameterizations were acceptable. We first computed $\hat{D}$ from the normalized variance and determined that non-uniqueness existed on the (Z, χ) manifold due to a nonzero value in $\mathcal{N}$ (and therefore $\hat{D}$ as well) at small bandwidths. We then compared the $\hat{D}$ curves for QoIs in the original 12-dimensional space with the two-dimensional projections and found that additional peaks were introduced through projection, indicating new scales of variation. Through subsampling the datasets, we assessed sensitivity of these new peak locations to data spacing and found that Range PCA, with a high sensitivity, also resulted in non-unique QoI representation. In comparison, Pareto PCA and (Z, ζ_{CO_2}) did not show such sensitivity and

were deemed acceptable parameterizations, with little difference in the remaining QoI peak locations (feature sizes). Visual inspection of Figure 1 supports this quantitative analysis.

4.2. Experimental Data

Next, we will demonstrate how our method can assess point-cloud datasets by analyzing parameterizations for an experimental combustion dataset. We consider the Sandia flame D dataset [36] which contains approximately 57,000 observations each for temperature and select species over six different heights in a CH_4/air piloted jet flame. Figure 14 shows the two-dimensional Pareto PCA and (Z, ζ_{CO_2}) representation for the Sandia flame D data colored by the main fuel species Y_{CH_4} . We can see that the data is less structured than the simulated flamelet data from Figure 1 and it is less clear if these parameterizations are adequate in two-dimensions. We will therefore be comparing parameterizations for Pareto PCA and Range PCA in two to four dimensions along with the two-dimensional (Z, ζ_{CO_2}) representation. Our analysis of QoIs will include experimental measurements of T , Y_{O_2} , $Y_{\text{H}_2\text{O}}$, Y_{CH_4} , Y_{CO_2} , and Y_{OH} . The Sandia flame D dataset does not contain observations of source terms and presents difficulties in extracting enough information to compute source terms since data for some species involved in the reactions of methane combustion are missing; therefore, source terms will be excluded from the QoIs in this analysis. We note that while the principal components are linear functions of all QoIs in this scenario, the QoIs become nonlinear functions of the principal components once a low-dimensional parameterization is chosen [20, 21]. After the dimension reduction, an inversion of the linear operator defined by PCA to obtain the original variables is only approximate. Therefore, while the QoIs considered for the experimental dataset may not exhibit the same degree of nonlinearity as source terms might, they can still be used in a meaningful demonstration of our quality assessment tool.

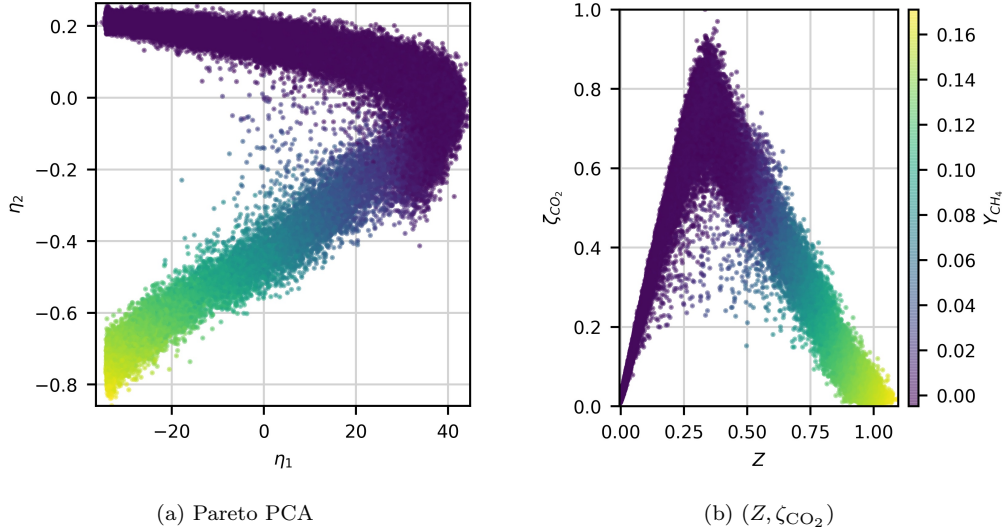


Figure 14. Two-dimensional parameterizations for the experimental Sandia flame D data colored by Y_{CH_4} .

Figure 15 shows the normalized variance and $\hat{D}$ over bandwidths for QoIs on the two-dimensional Pareto PCA and (Z, ζ_{CO_2}) projected manifolds as well as the original ten-dimensional space. The original ten-dimensions include temperature and all

select species contained in the dataset. We can see that the curves for the normalized variance are still smooth and the analysis applies to our point-cloud dataset even in ten-dimensions. The radical species Y_{OH} is showing the most difference between the original space and the projection, just as seen with the simulated data. In the original ten-dimensional space, all QoIs are showing a single feature size, while in the two-dimensional projections, additional scales of variation are being introduced (mainly for Y_{OH}). Therefore, we need to determine whether or not these new feature sizes can be attributed to non-uniqueness.

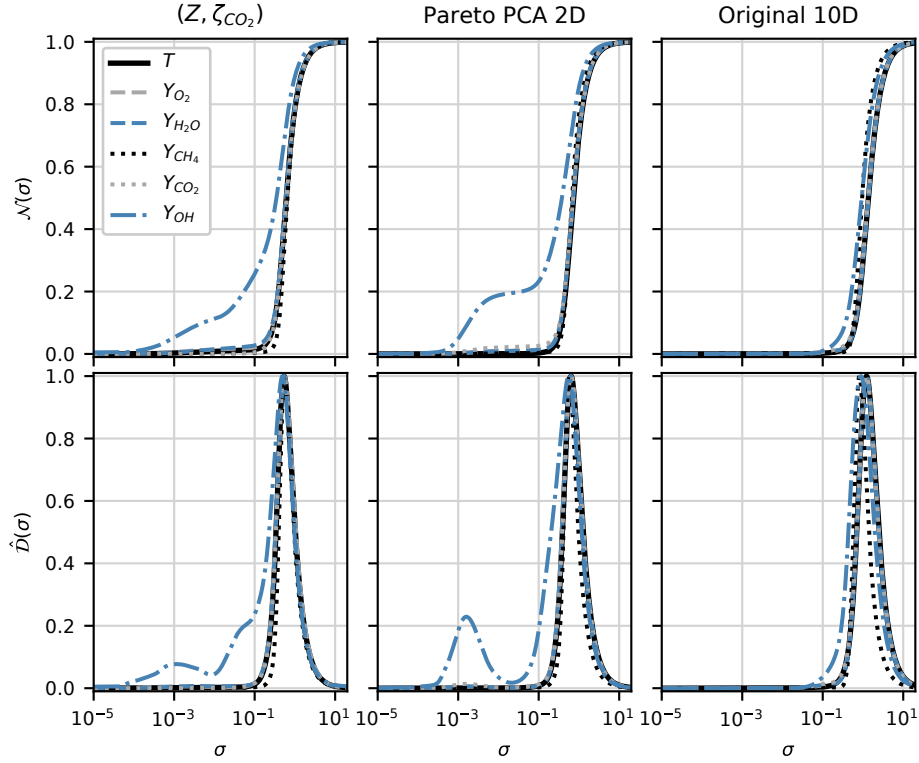


Figure 15. Normalized variance analysis for sample two-dimensional projections of the experimental Sandia flame D data along with the original ten-dimensional space.

We can also examine the peaks of $\hat{\mathcal{D}}$ for higher dimensional parameterizations. Figure 16 displays $\hat{\mathcal{D}}$ for the QoIs on the four-dimensional manifold projections. We find that Pareto PCA in four dimensions, shown in Figure 16a, still shows additional scales of variation not present in the original space shown in Figure 15. However, in four dimensions, Range PCA, shown in Figure 16b, shows a single dominant feature size for all QoIs just as in the original space. Therefore, the Range PCA 4D projection is an adequate representation of the QoIs for the experimental dataset. We now need to determine whether or not the other parameterizations are acceptable.

In order to determine if new scales of variation for QoIs on the low-dimensional manifold projections are due to non-uniqueness, we again use data subsampling to analyze sensitivity of these scales to data spacing. Figure 17 shows how peaks in $\hat{\mathcal{D}}$ for Y_{OH} (which showed the most noticeable difference compared to the original high-dimensional space) shift due to different subsampling percentages. The reported shifts

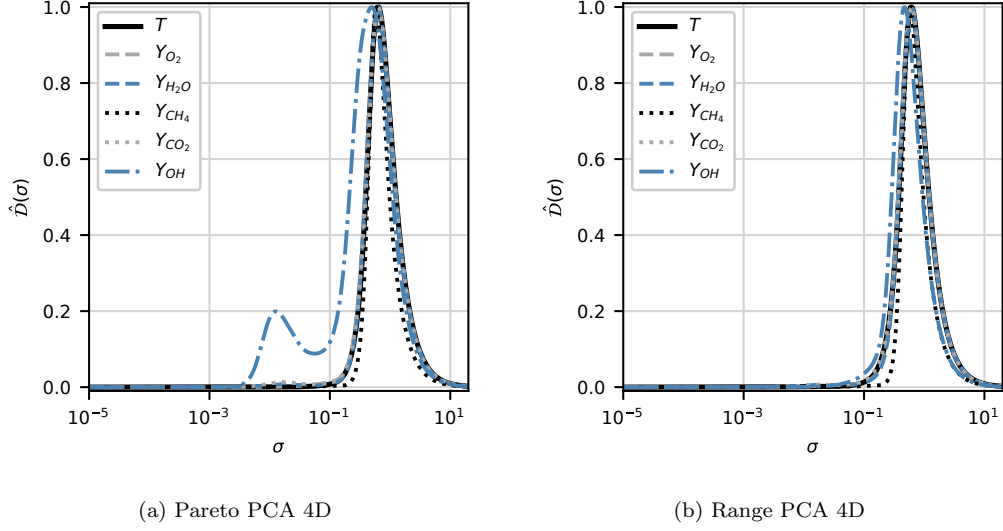


Figure 16. $\hat{D}(\sigma)$ for four dimensional PCA parameterizations of the Sandia flame D data.

in peak locations on a logarithmic scale relative to 100% sampling come from averaging $\hat{D}$ over five realizations of a randomly selected sample based on a percentage of the projected manifold dataset. We find that the tallest peak, represented with solid lines, does not experience a significant shift (3-5% on a logarithmic scale) with subsampling percentage. In comparison, secondary peaks, represented with dashed lines, show a 11-20% shift in $\log_{10}(\sigma_{\text{peak}})$ as samples are decreased to 25% of the projected dataset. Therefore, all of the manifold parameterizations considered here, excluding Range PCA 4D, show signs of non-uniqueness and are not acceptable parameterizations of the Sandia flame D dataset. We note that a dashed line for the Range PCA 4D manifold is not present in Figure 17 because it does not have a secondary peak for Y_{OH} .

We find that the Y_{OH} QoI shows scales of variation indicative of non-uniqueness in the parameterizations, excluding Range PCA in four dimensions. However, the lower-dimensional parameterizations show promise in representing the other QoIs. This demonstrates how non-uniqueness is QoI-dependent and how our manifold assessment technique can determine when scales can be attributed to non-uniqueness injected by a parameterization. Based on the analysis for the given set of QoIs, the Range PCA projection into four dimensions is the only acceptable projection. Further analysis on the QoI feature sizes could be done to try differentiating the remaining manifolds if Y_{OH} was not included as a QoI. This also demonstrates how the required dimensionality of a parameterization can be QoI-dependent and can be determined by analyzing QoI-uniqueness over manifolds of various dimensionality. We have shown that our quantitative manifold assessment technique can compare parameterizations of a specific dimensionality as well as parameterizations across dimensions. It is not limited to a particular method of parameterization, dimensionality, or data structure.

5. Conclusions

Dimension reduction is commonly used to facilitate large-scale simulation of high-dimensional dynamical systems such as combustion. Surrogate models must then accurately represent QoIs over the low-dimensional parameterization in order for simula-

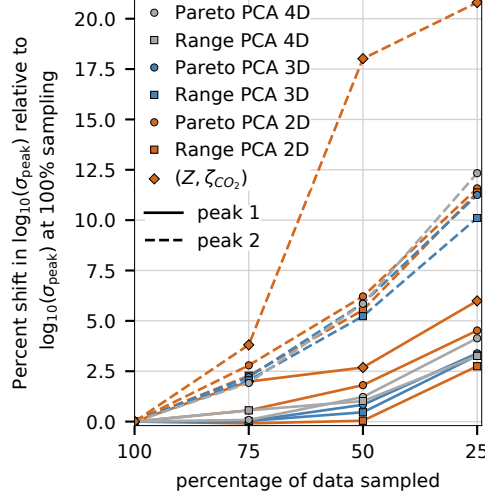


Figure 17. Percent shift in logarithmic peak locations for $\hat{\mathcal{D}}(\sigma)$ of Y_{OH} for various parameterizations of the experimental Sandia flame D data. Peak locations (σ_{peak}) are determined after averaging five realizations of $\hat{\mathcal{D}}(\sigma)$ from randomly selected samples of the manifold projection datasets and the shift is reported relative to $\log_{10}(\sigma_{\text{peak}})$ with 100% sampling. “peak 1” (solid lines) refers to the tallest peak and “peak 2” (dashed lines) refers to the second tallest peak (if it exists). Note that Range PCA 4D does not have a dashed line since it does not create a second peak for Y_{OH} as shown in Figure 16b.

tions to succeed. This suggests manifold quality assessments should take into account features that may pose difficulties for reconstruction, such as non-uniqueness in QoIs introduced through a parameterization. Existing methods, however, limit analysis to the independent variables and low-dimensional representation.

We have proposed a technique that quantitatively assesses manifold quality through characterizing feature sizes of QoIs using a normalized variance measure over varying bandwidth/filter width. Locations of peaks in a scaled derivative of the normalized variance over logarithmically scaled bandwidths represent fast increases in QoI-variation with increasing bandwidth and illustrate feature sizes. These peaks can be used to determine when parameterizations introduce new scales of variation or feature sizes compared to the original high-dimensional space through the introduction of new peaks. Furthermore, quantifying the sensitivity of these peaks to data spacing using subsampling can help distinguish between non-uniqueness and a characteristic feature size.

The outlined analysis allows for comparing parameterizations of a given dataset for representing the QoIs in arbitrary dimensions. Larger feature sizes are more desirable and detection of non-uniqueness in QoI representation means a projection is unacceptable. The proposed manifold quality assessment technique has been successfully demonstrated on smooth simulated and noisy experimental datasets from combustion for comparing parameterizations from physics-based and data-based reduction methods. This technique is not limited to a specific reduction method, dimensionality, or data structure; and it is applicable to scenarios demanding a QoI-dependent assessment such as modeling requirements in reduction of high-dimensional dynamical systems.

Acknowledgments

The authors gratefully acknowledge the support of Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

References

- [1] I. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York, 2002.
- [2] S.T. Roweis and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science 290 (2000), pp. 2323–2326.
- [3] J.B. Tenenbaum, V.d. Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science 290 (2000), pp. 2319–2323.
- [4] Z. Zhang and H. Zha, *Principal manifolds and nonlinear dimensionality reduction via tangent space alignment*, SIAM Journal on Scientific Computing 26 (2004), pp. 313–338.
- [5] K.Q. Weinberger and L.K. Saul, *Unsupervised learning of image manifolds by semidefinite programming*, International Journal of Computer Vision 70 (2006), pp. 77 – 90.
- [6] S.B. Pope, *Small scales, many species and the manifold challenges of turbulent combustion*, Proceedings of the Combustion Institute 34 (2013), pp. 1 – 31.
- [7] L. Chen and A. Buja, *Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis*, Journal of the American Statistical Association 104 (2009), pp. 209–219.
- [8] J. Venna and S. Kaski, *Local multidimensional scaling*, Neural Networks 19 (2006), pp. 889 – 899.
- [9] J.A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, Springer, 2007.
- [10] P. Zhang, Y. Ren, and B. Zhang, *A new embedding quality assessment method for manifold learning*, Neurocomputing 97 (2012), pp. 251 – 266.
- [11] D. Chen, J. Lv, J. Yin, H. Zhang, and X. Li, *Angle-based embedding quality assessment method for manifold learning*, Neural Computing and Applications 31 (2019), pp. 839–849.
- [12] J. Liang, S. Chenouri, and C.G. Small, *A new method for performance analysis in nonlinear dimensionality reduction*, Statistical Analysis and Data Mining: The ASA Data Science Journal 13 (2020), pp. 98–108.
- [13] N. Peters, *Laminar diffusion flamelet models in non-premixed turbulent combustion*, Progress in Energy and Combustion Science 10 (1984), pp. 319 – 339.
- [14] J.A. van Oijen and L.P.H. de Goey, *Modelling of premixed laminar flames using flamelet-generated manifolds*, Combustion Science and Technology 161 (2000), pp. 113–137.
- [15] O. Gicquel, N. Darabiha, and D. Thévenin, *Laminar premixed hydrogen/air counterflow flame simulations using flame prolongation of ILDM with differential diffusion*, Proceedings of the Combustion Institute 28 (2000), pp. 1901–1908.
- [16] G. Bansal, A. Mascarenhas, and J.H. Chen, *Identification of intrinsic low dimensional manifolds in turbulent combustion using an isomap based technique*, Seventh US National Meeting of the Combustion Institute (2011).
- [17] C.E. Frouzakis, Y.G. Kevrekidis, J. Lee, K. Boulouchos, and A.A. Alonso, *Proper orthogonal decomposition of direct numerical simulation data: Data reduction and observer construction*, Proceedings of the Combustion Institute 28 (2000), pp. 75 – 81.
- [18] S.J. Danby and T. Echekki, *Proper orthogonal decomposition analysis of autoignition simulation data of nonhomogeneous hydrogen-air mixtures*, Combustion and Flame 144 (2006), pp. 126 – 138.
- [19] J.C. Sutherland and A. Parente, *Combustion modeling using principal component analysis*, Proceedings of the Combustion Institute 32 (2009), pp. 1563 – 1570.

- [20] A. Biglari and J.C. Sutherland, *A filter-independent model identification technique for turbulent combustion modeling*, Combustion and Flame 159 (2012), pp. 1960 – 1970.
- [21] Y. Yang, S.B. Pope, and J.H. Chen, *Empirical low-dimensional manifolds in composition space*, Combustion and Flame 160 (2013), pp. 1967–1980.
- [22] B.J. Isaac, J.N. Thornock, J. Sutherland, P.J. Smith, and A. Parente, *Advanced regression methods for combustion modelling using principal components*, Combustion and Flame 162 (2015), pp. 2592 – 2601.
- [23] T. Echehki and H. Mirgolbabaei, *Principal component transport in turbulent combustion: A posteriori analysis*, Combustion and Flame 162 (2015), pp. 1919 – 1933.
- [24] R. Ranade and T. Echehki, *A framework for data-based turbulent combustion closure: A posteriori validation*, Combustion and Flame 210 (2019), pp. 279 – 291.
- [25] M.R. Malik, P. Obando Vega, A. Coussement, and A. Parente, *Combustion modeling using principal component analysis: A posteriori validation on sandia flames d, e and f*, Proceedings of the Combustion Institute (2020).
- [26] A. Parente, J.C. Sutherland, L. Tognotti, and P.J. Smith, *Identification of low-dimensional manifolds in turbulent flames*, Proceedings of the Combustion Institute 32 (2009), pp. 1579 – 1586.
- [27] A. Parente, J. Sutherland, B. Dally, L. Tognotti, and P. Smith, *Investigation of the mild combustion regime via principal component analysis*, Proceedings of the Combustion Institute 33 (2011), pp. 3333 – 3341.
- [28] A. Coussement, O. Gicquel, and A. Parente, *Kernel density weighted principal component analysis of combustion processes*, Combustion and Flame 159 (2012), pp. 2844 – 2855.
- [29] H. Mirgolbabaei, T. Echehki, and N. Smaoui, *A nonlinear principal component analysis approach for turbulent combustion composition space*, International Journal of Hydrogen Energy 39 (2014), pp. 4622 – 4633.
- [30] H. Mirgolbabaei and T. Echehki, *Nonlinear reduction of combustion composition space with kernel principal component analysis*, Combustion and Flame 161 (2014), pp. 118 – 126.
- [31] I. Noda, *Scaling techniques to enhance two-dimensional correlation spectra*, Journal of Molecular Structure 883-884 (2008), pp. 216 – 227, progress in two-dimensional correlation spectroscopy.
- [32] A. Parente and J.C. Sutherland, *Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity*, Combustion and Flame 160 (2013), pp. 340 – 350.
- [33] U. Maas and D. Thévenin, *Correlation analysis of direct numerical simulation data of turbulent non-premixed flames*, Symposium (International) on Combustion 27 (1998), pp. 1183–1189.
- [34] J.C. Sutherland, P.J. Smith, and J.H. Chen, *A quantitative method for a priori evaluation of combustion reaction models*, Combustion Theory and Modelling 11 (2007), pp. 287–303.
- [35] W. Härdle, *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press, 1990.
- [36] R.S. Barlow and J.H. Frank, *Effects of turbulence on species mass fractions in methane/air jet flames*, Symposium (International) on Combustion 27 (1998), pp. 1087 – 1095.