Interaction-Aware Sensitivity Analysis for Aerodynamic Optimization Results using Information Theory

Patricia Wollstadt, Sebastian Schmitt

2021

Preprint:

Interaction-Aware Sensitivity Analysis for Aerodynamic Optimization Results using Information Theory

1st Patricia Wollstadt
Honda Research Institute Europe GmbH
Offenbach/Main, Germany
patricia.wollstadt@honda-ri.de, 0000-0002-7105-5207

2nd Sebastian Schmitt
Honda Research Institute Europe GmbH
Offenbach/Main, Germany
sebastian.schmitt@honda-ri.de, 0000-0001-7130-5483

***Abstract*—An important issue during an engineering design process is to develop an understanding which design parameters have the most influence on the performance. Especially in the context of optimization approaches this knowledge is crucial in order to realize an efficient design process and achieve high-performing results. Information theory provides powerful tools to investigate these relationships because measures are model-free and thus also capture non-linear relationships, while requiring only minimal assumptions on the input data. We therefore propose to use recently introduced information-theoretic methods and estimation algorithms to find the most influential input parameters in optimization results. The proposed methods are in particular able to account for interactions between parameters, which are often neglected but may lead to redundant or synergistic contributions of multiple parameters. We demonstrate the application of these methods on optimization data from aerospace engineering, where we first identify the most relevant optimization parameters using a recently introduced information-theoretic feature-selection algorithm that accounts for interactions between parameters. Second, we use the novel partial information decomposition (PID) framework that allows to quantify redundant and synergistic contributions between selected parameters with respect to the optimization outcome to identify parameter interactions. We thus demonstrate the power of novel information-theoretic approaches in identifying relevant parameters in optimization runs and highlight how these methods avoid the selection of redundant parameters, while detecting interactions that result in synergistic contributions of multiple parameters.**

***Index Terms*—feature selection, information theory, partial information decomposition, aerospace design optimization, engineering data mining**

I. Introduction

Optimizing the performance of systems of parts is a central task during an engineering design process. For example, in automotive or aerospace engineering, the shape of individual parts is commonly optimized to improve aerodynamic performance using computer aided design (CAD) methods. Typically, engineers wish to understand which changes in a shape, carried out during the optimization, lead to the improved behavior. Hereby, it is often of interest to account for interactions between parameters such as to identify which parameters influence a shape's fitness only when considered jointly [1], [2]. We therefore present a novel, information-theoretic approach for the identification of optimization parameters most relevant to changes in a shape's fitness, which accounts for interactions between parameters with respect to the fitness, such as to identify parameters that interact jointly with the target. We further utilize recently introduced information-theoretic measures to quantify interactions between features. We demonstrate the applicability of our approach on a set of realistic turbofan rotor blade optimization runs [3], but strongly believe that it is of interest for a wide range of engineering design optimization scenarios.

Information theory [4] is a powerful tool for the analysis of dependencies between variables. Information-theoretic methods, such as the mutual information (MI), are model-free and are able to capture dependencies of arbitrary order, while requiring only minimal assumptions about the data for their estimation when using state-of-the-art estimators [5]. These properties make information-theoretic measures particularly promising tools for the analysis of data in the engineering domain [6], for example, results from optimization runs [7]. Here, the relationship between parameters and the optimization objective is expected to be highly non-linear and the number of data samples is typically rather limited because the evaluation of fitness functions is costly. Furthermore, data distributions are typically not known and are expected to be highly biased due to the fact that data are generated by an optimization algorithm. As a result, high-quality global surrogate models that cover substantial parts of the search domain are most likely not available to understand optimization runs [3]. Thus, there is a need for methods that allow for a post-hoc analysis of optimization parameters and their influence on the optimization outcome.

We use a recently introduced algorithm for inferring relationships between variables that uses a conditional mutual information criterion (CMI) as a selection criterion [8], [9]. Using the CMI for selecting variables allows to account for interactions between variables such as redundancies, but also synergistic contributions [10]. Furthermore, we use the recently introduced partial information decomposition (PID) framework to investigate selected variables for interactions

with respect to the target variable. We apply our approach to data from realistic turbofan blade aerodynamic optimization runs that use computational fluid dynamics (CFD) to evaluate a shape's fitness [3]. We propose a parametrization of the turbofan blade geometry that allows for application of the proposed algorithm and compare our algorithm's performance to related information-theoretic feature selection criteria. To our knowledge, this work is the first using PID for sensitivity analysis in aerodynamic optimization data.

II. Methods

A. Optimization and Simulation Setup

We use data from realistic optimization runs on turbofan rotor blade geometries that was previously described and published in [3]. For details on the data generation process refer to the original publication. Fig. 1A shows a schematic of a turbojet engine and the corresponding turbofan rotor blade geometry.

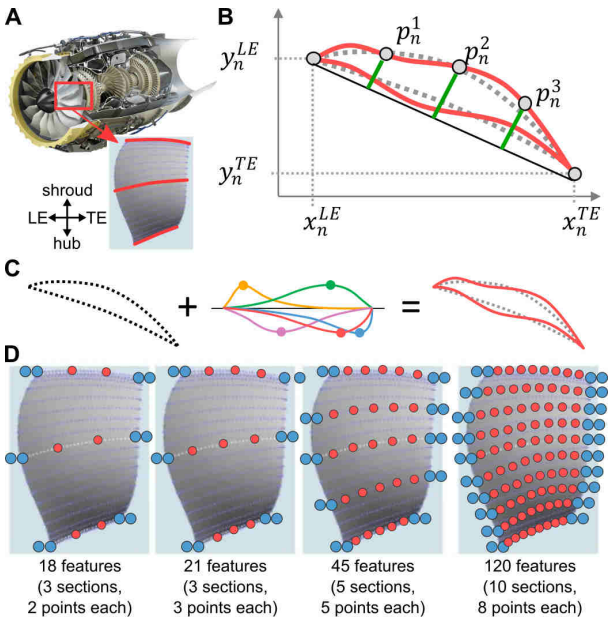

Fig. 1. Turbofan blade, modification and feature sets. **A** Investigated turbofan blade and Honda HF 120 jet engine, orange lines indicate cross-sections at which shape modifications were performed (LE: leading edge, TE: trailing edge). **B** Features selected form one blade section. **C** Modification of one blade section through addition of Hicks-Henne functions used during the shape optimization. **D** Location of the extracted feature sets, defined by varying numbers of sectional cuts through the geometry and number of points per section (red markers). Blue markers indicate leading and trailing edge features, each comprising two features for the x- and y-coordinate of the edge, respectively.

A rotor blade is optimized by starting with a baseline-shape that is modified under the objective of minimizing a target function. The shape is modified by deforming three cross-sections of the shape, where each section is a cylindrical cut of the geometry. We consider one section at the hub, one at mid-span height, and one at the shroud of the blade as indicated by the red lines in the inset of Fig. 1A.

Each section is deformed independently by the following manipulations: (i) rotation of the section around the leading edge (LE) point, (ii) movement of the section in the axial-meridional plane, and (iii) deformation of the section profile by adding Hicks-Henne shape functions [11], which is a common approach in 2D airfoil design and is illustrated in Fig. 1C. The Hicks-Henne function is defined as

$$b(x, x_0) = \left[\sin\left(\pi x^{\frac{\log(0.5)}{\log(x_0)}}\right)\right]^2, \tag{1}$$

where $x \in [0, 1]$ parametrizes the chord length of each section and x_0 is the location of the maximum of each shape function. We placed the maxima of N_{HH} shape functions per section at equally spaced locations along the cord length, $x_0(i) = \frac{i}{N_{HH}+1}$ where $i = 1, \ldots, N_{HH}$. Considering all three possible manipulations, section rotation, movement, and deformation with Hicks-Henne functions, the total number of free shape parameters is $N = 3\,(N_{HH} + 3)$.

For the optimization of shape parameters, we used a covariance matrix adaptation evolutionary strategy (CMA-ES) [12] with a population size of $\lambda = 12$ and $\mu = 4$ parents which we ran for 161 generations, which amounts to 1932 evaluations, i.e., data samples, per run. We used an initial step size of $\sigma = 0.05$ in relative units of the maximal allowed variation (i.e., a $5\,\%$ initial variation). We performed four optimization runs, where two runs were performed with $N_{HH} = 3$ and two runs with $N_{HH} = 7$, which lead to 18 and 30 free parameters to be determined by the optimization, respectively. Each run was initialized using a different random seed. These parameter settings are derived from best practices which try to balance the exploration and exploitation capabilities of each optimization run, to arrive at manageable optimization runtimes (each CFD simulation of a blade takes about 2 hours on 32 cores), and utilize the HPC infrastructure most efficiently.

The optimization target was to maximize the aerodynamic efficiency of the rotor blade at cruising conditions, which is estimated by calculating the isentropic efficiency of the blade,

$$\eta = \frac{\left(\frac{P_{T,outlet}}{P_{T,inlet}}\right)^{\frac{\gamma-1}{\gamma}} - 1}{\frac{T_{T,outlet}}{T_{T,inlet}} - 1}, \tag{2}$$

where P_T and T_T are the mass-flow averaged total pressure and total temperature at the specified location and $\gamma = 1.4$ is the heat capacity ratio (see, for example, [13]).

The boundary conditions of the CFD simulation mimic the behavior of a jet engine under cruising conditions. Each blade was evaluated with a CFD simulation which employed the compressible flow solver `steadyCompressibleMRFFoam` from the OpenFOAM CFD suite (version `foam-extend-3.2`), adapted to be more robust for trans-sonic simulations [14]. The fitness of a blade was calculated as

$$f = 1 - \eta_{avg} + P, \tag{3}$$

where η_{avg} denotes the isentropic efficiency of the blade of Eq. (2), averaged over the last 100 iterations of the solver. P represents a penalty term that increases and thus worsens the fitness if the CFD simulation does not show good convergence

or if the generated blade geometry is not feasible. See the original publication [3] for more details on the simulation setup, the optimization and the data generation. The fitness values during the optimization runs as function of the generations is shown in Fig. 2.

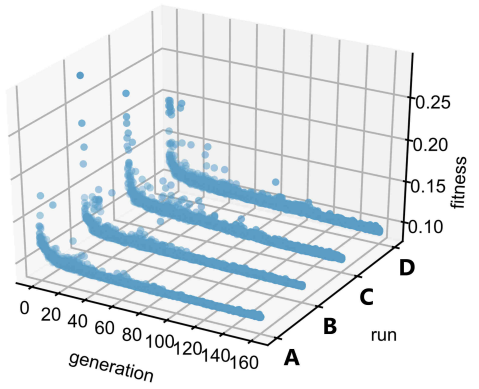

Fig. 2. Raw fitness values over generations for four investigated optimizations.

B. Feature Extraction of Turbofan Blade Geometries for Sensitivity Analysis

We ran four optimizations with varying numbers of parameters for shape modification. In a next step, we wished to identify the locations at which modifications were most relevant to a blade's fitness. To apply the proposed information-theoretic approach, we first had to find suitable features that represented the blade geometry's surface and could be used as input features for the algorithm (e.g., [15]). To this end, we considered multiple sectional cuts through the blade geometry. At each sectional cut n, we placed N_{points} points equally spaced along the chord line and recorded the absolute distance from the actual blade surface to the chord line at these locations (Fig. 1B). Furthermore, we considered the x- and y-coordinates of the leading edge (LE), x_n^{LE} and y_n^{LE}, as well as the x- and y-coordinates of the trailing edge (TE), x_n^{TE} and y_n^{TE}. We varied the number of points and sections used for the features to investigate the stability of results over various representations of the geometry. We used 3 sectional cuts with 2 points, resulting in 18 features, 3 cuts with 3 points, resulting in 21 features, 5 cuts and 5 points resulting in 45 features, and 10 cuts and 8 points, resulting in 120 features (Fig. 1D).

C. Information-Theoretic Preliminaries

Before introducing the algorithm used to identify the most relevant locations of modification, we introduce the necessary information-theoretic preliminaries (for a more detailed introduction see [16]).

The algorithm uses a conditional mutual information (CMI) to quantify the influence a single feature has on the fitness, in the context of further features. The CMI is defined as

$$I(X;Y|Z) = \sum_{x\in\mathcal{A}_X, y\in\mathcal{A}_Y, z\in\mathcal{A}_Z} p(x,y,z)\log\frac{p(x|y,z)}{p(x|z)}, \quad (4)$$

where X, Y, Z are random variables with realizations x, y, z, and $p(x)$ is a shorthand for the probability distribution $p(X=x)$. The CMI quantifies the average information that X has about Y, given the outcome of Z is known. The CMI is symmetric in X and Y, and $I(X;Y|Z)\geq 0$. Further, each random variable may also be replaced by a set of variables, e.g., $\mathbf{X}$, and thus quantifying the information a set of variables provides about a second variable, Y or set of variables, $\mathbf{Y}$.

Note that conditioning the information X provides about Y on a third variable, Z, $I(X:Y|Z)$ may have two effects: first, information that is *redundantly* present in both X and Z about Y is removed from the information X alone provides about Y (as measured by the unconditioned MI, $I(X;Y)$). Second, information that is provided *synergistically* by X and Z together about Y is added to the information X alone is providing about Y [17]. Hence, the CMI quantifies the information X provides *uniquely* about Y and the information X and Z provide jointly about Y in a *synergistic* fashion; at the same time, *redundant* contributions in X and Z about Y are excluded. See also [10] for a discussion of the use of the CMI for feature selection.

As an example of synergistic information contribution, consider a binary `xor`-gate with iid. inputs, X and Z, and output Y. Inputs X and Z alone, each provide no information about the output Y, such that $I(X;Y)=I(Z;Y)=0$. Only by conditioning on the second input, the information the first input provides is "decoded" and $I(X;Y|Z)=I(Z;Y|X)=1$. Here, the two inputs provide information about the output in an exclusively synergistic fashion.

The framework to decompose the information two variables contribute about a third into unique, redundant, and synergistic contributions has only recently been introduced and is termed *Partial Information Decomposition* (PID) [17] (Fig. 3A, see also [18], [19]). PID extends classical information theory by providing axioms that allow to decompose the joint information two input variables X and Z provide about a target variable Y, $I(Y;X,Z)$, into the information provided uniquely by each X and Y, information provided redundantly by X and Y, and information provided synergistically when considering X and Y jointly. Note that such a detailed decomposition of the information contributed by two variables about a third was not possible using existing information-theoretic concepts, e.g., the (C)MI or Shannon entropy, as shown by Williams and Beer [17] and illustrated in Fig. 3B.

In the present work, we use the PID framework to identify interactions between features with respect to the blade's fitness. In particular, we estimate the synergistic information contribution of features and sets of features to identify those feature combinations that provide information about the fitness primarily when considered jointly.

D. Identification of Most Relevant Features using Information-Theoretic Feature Selection Algorithm

We used a recently introduced forward-selection algorithm for feature selection [8]–[10] to identify the most relevant blade features with respect to the optimization outcome. The

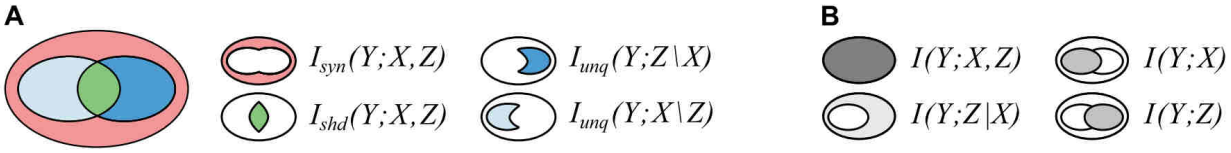

Fig. 3. **A** Partial information decomposition diagram: decomposition of the joint mutual information, $I(Y;X,Z)$ into unique information of each input variable (light and dark blue), redundant information (green), and synergistic information (red). **B** Corresponding, classical information-theoretic terms.

algorithm uses a CMI criterion for iterative feature selection, which measures the MI between a feature to be selected and the fitness, conditional on all already selected features. Thus, the CMI criterion, includes features not only based on their individual (unique) information contribution to the fitness, but also accounts for synergistic effects between the currently considered feature and the already selected feature set. Lastly, the inclusion criterion ensures that redundancies between features are avoided. For a detailed discussion of the algorithm and the CMI as a feature-selection criterion see [10]. See algorithm 1.

Algorithm 1 Forward feature selection

```
function SELECTFEATURES(X, Y, α_crit)
    S ← ∅                                   ▷ Initialization of feature set
    while X ≠ ∅ do                          ▷ Find next candidate feature
        F ← max_{X∈X} I(X;Y|S)
        α ← permutationtest(I(F;Y|S))
        if α < α_crit then                  ▷ Contribution is significant
            S ← S ∪ F                       ▷ Add candidate to feature set
            X ← X \ F
        else                                ▷ Contribution is not significant
            break                           ▷ Terminate inclusion
    return S                                ▷ Final feature set
```

The algorithm starts with an empty feature set $\mathbf{S} = \emptyset$, the set of all input variables, $\mathbf{X}_0 = \mathbf{X}$, and the target variable Y. Features are selected iteratively, where in each iteration, i, the algorithm selects the feature that maximizes the criterion,

$$F_i = \max_{X \in \mathbf{X}_i} I(X;Y|\mathbf{S}_i), \tag{5}$$

where $\mathbf{X}_i \subseteq \mathbf{X}$ denotes the remaining input variables in iteration i, and $\mathbf{S}_i$ the set of already selected features. The identified maximum contribution is tested for statistical significance using non-parametric permutation testing and a testing scheme that controls the family-wise error rate (see [9] for a detailed description of the test). If the information contributed by F_i as measured by the CMI is statistically significant, F_i is included in the set of selected features, $\mathbf{S}_i$ and removed from the set of remaining variables, $\mathbf{X}_i$,

$$\begin{aligned} \mathbf{S}_{i+1} &= \mathbf{S}_i \cup F_i\,, \\ \mathbf{X}_{i+1} &= \mathbf{X}_i \setminus F_i\,. \end{aligned} \tag{6}$$

Note that statistical testing of the CMI estimate is necessary because while in theory the CMI is zero for (conditionally) independent variables, this may not be the case when estimating the CMI from finite data, due to the known bias of information-theoretic estimators (e.g., [20]). Instead, the test evaluates whether the estimate significantly differs from the distribution of estimates from permuted data and thus tests the Null-hypothesis of no dependence between the feature and the target in the context of the already selected feature set. The statistical test not only handles the estimation bias, but also provides an automatic stopping criterion for feature selection, because the algorithms stops if no remaining variable provides significant information about the target, given the already selected feature set. The number of features included in the selected feature set can indirectly be influenced by changing the critical alpha-level, α_{crit}, of the statistical test, i.e., the threshold an individual test in iteration i has to pass to allow for inclusion of candidate feature F_i. We here used $a_{crit} = 0.05$, where lowering α_{crit} leads to a more strict criterion and thus to the selection of fewer features in general, and vice versa.

For practical estimation, we use an implementation of the algorithm as part of the IDTxl python toolbox [8]–[10], which uses a k-nearest-neighbor-based estimator for MI and CMI estimation from continuous data [5], which—while not being bias-free—has shown to provide the most favorable bias properties compared to other approaches [5], [21], [22].

E. Post-hoc Analysis of Feature Interactions by Estimating Synergistic Information Contribution

After selecting the most relevant geometric features for each optimization run using the presented forward-selection algorithm, we identify interactions between features with respect to the fitness by estimating the synergy between all pairs of selected features and the fitness. We use a PID estimator introduced in [23], also implemented in the IDTxl toolbox [8].

III. Results

A. Identified Features and Interactions Between Features

The locations of features for the four optimization runs and the four extracted feature sets of the blade surface are shown in Fig. 4. Here, the first two markers in each row indicate the x- and y-coordinates of the leading edge, x_n^{LE}, y_n^{LE}, while the last two markers indicate the coordinates of the trailing edge, x_n^{TE}, y_n^{TE} (both are in blue). The bottom row indicates the section closest to the hub, while the top row indicates the section closest to the shroud. Markers between the first and last two markers in each row indicate geometric features from left to right, p_n^m, where $n \in \{1,\ldots,N\}$ indicates the section

number from hub to shroud and $m \in \{1, \ldots, M\}$ indicates the feature index. Hence, the total number of input variables per feature set was $N_{feat} = NM + 4N$. Panels A and B, and panels C and D each show optimization runs with identical setup but different random initialization for $N_{HH} = 3$ (A and B) and for $N_{HH} = 7$ (C and D).

Colored markers indicate relevant features identified by the algorithm. Dashed lines indicate the three pairs of features with highest synergy over all feature pairs.

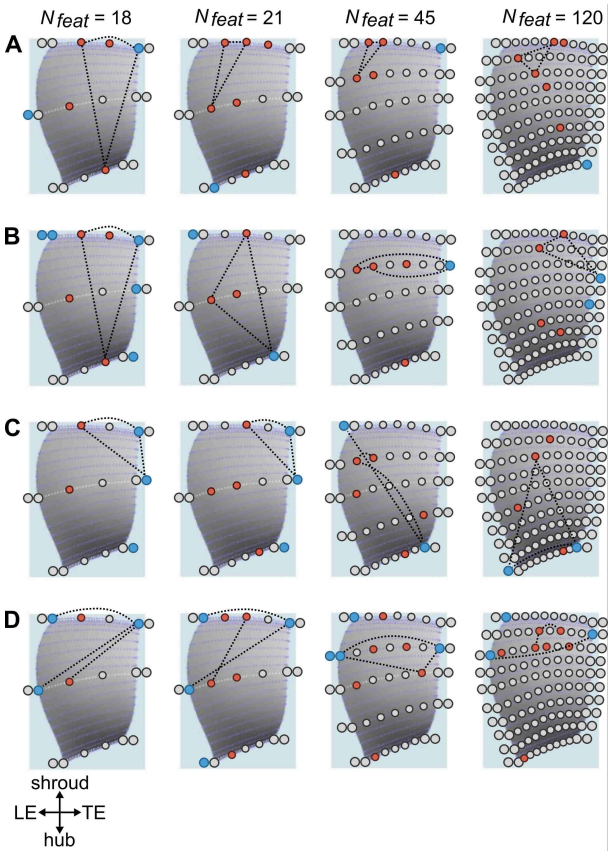

Fig. 4. Locations of selected features and identified interactions for four runs (rows A to D). Each column shows a different feature set, using 18, 21, 45, or 120 features respectively. Colored markers indicate selected features, dotted lines indicate the three pairs with the highest interaction wrt. the blade's fitness as measured by the synergistic information. The meaning of the colors is the same as in Fig. 1D.

We first note that the selected features are not completely consistent between runs which is expected. The data for each case was generated by an optimization run which is a highly structured process, and therefore, the feature space is sampled very inhomogeneously. Additionally, the blade regions with the largest deformations differ between runs [3], leading to variations in the extracted features. However, there are regions which are identified to be important in all runs, for example, at around 30 % chord length from the LE in the region from mid-span to blade tip (i.e. the upper forward region). This region is expected to have high influence due to the shock-system built-up [24]. Similarly, the region near the TE, and in particular close to the tip, is directly influencing the exit-flow angle and thus affecting the efficiency strongly. The location at the hub is also consistently identified as important, but the exact location along the chord line varies between the runs.

Comparing the selected features of each run between the different feature sets provides a consistent picture for the smaller feature sets $N_{feat} = 18$, 21, and 45. The apparent differences can be understood by considering the peculiarities of the data and the PID-based selection method. First of all, it is expected that for each feature set strong correlations and redundancy are present in the features, due to the deformation method used to generate the blades. Only three sections (at hub, mid-span and shroud) were allowed to change independently and the changes were linearly interpolated in-between, leading to many features being linear combinations of others. In addition, the Hicks-Henne-based deformations of each section also induces smooth changes with possibly highly correlated and thus potentially redundant neighboring features. Also, the optimization algorithm induces correlated changes of parameters, i.e., blade regions, once it starts to converge to some (local) optimum. Therefore, the features from the feature set with $N_{feat} = 21$ selected on the mid-span section are replaced by (Fig. 4A and B) or augmented with (C and D) more informative features on the second section from the tip.

For $N_{feat} = 45$, high values of the redundancy are observed between the selected features and the not selected features which are close to the locations of the selected features form the smaller sets (not shown).

For the largest feature set with $N_{feat} = 120$ the selected features are consistent with the smaller feature sets in the above described manner for the case D, but are only partially consistent or even seem inconsistent for the other cases A, B and C. This is understandable from the insights described above. Extracting 120 features from designs which are created with only 18 (A and B) or 30 (C and D) independent parameters constitutes a vast over-parametrization of the independent influence factors, and results in huge redundancy in the feature set. In that case, the selected features are strongly influenced by the statistical variations of the rather few and highly structured 1932 data samples. Multiple sets of features could be selected which would be almost equally informative regarding the fitness, but which have different distribution of selected features over the blade region. Which set will be finally selected is strongly influenced by its ability to describe the statistical fluctuations of the data set. From the theoretical perspective this is correct, as the selected features represent the most informative features with respect to the fitness values *for the given data set*. However, the value to the engineer might be limited, as the most informative set does not necessarily represent the most important engineering design changes.

B. Prediction of Optimization Results

To validate the identified set of relevant parameters for each combination of number of features and optimization run, we used the selected features to predict the fitness values of each blade across the optimization run. We compared the features selected by our algorithm to features selected by

the FEAST toolbox [25] and features selected by standard machine learning approaches (linear Pearson correlation, MI, decision trees, extra trees, random forest, LARS).

The FEAST toolbox implements a variety of information-theoretic feature selection criteria based on the MI and applies them to rank features. These criteria do not consider interactions between features, i.e., features are evaluated solely based on their *individual* contribution to the target. Hence, synergistic effects as well as redundancies are not accounted for (see also [10] for a comparison of the selection criteria to the regular CMI). Also, the toolbox does neither provide means to handle estimator bias nor an automatic stopping criterion for feature inclusion. As the toolbox only handles discrete variables, we binned the data prior to feature selection.

We used the following selection criteria implemented in FEAST: Joint MI (JMI) [26], MI Maximization (MIM) [27], Max-Relevance Min-Redundancy (MRMR) [28], Conditional MI Maximization (CMIM) [29], Double Input Symmetrical Relevance (DISR) [30], Conditional Infomax Feature Extraction (CIFE) [31], Interaction Capping (ICAP) [32], Conditional Redundancy [25], Relief [33], and the CMI estimated from binned data. We predicted the fitness from the different selected feature sets using k-nearest-neighbor regression with number of neighbors, $k = 1$. Since the FEAST toolbox does not provide a stopping criterion, but just ranks the features by importance, we performed predictions from feature sets up to a size of 10 features, which was the maximum feature set size identified by our algorithm through statistical testing.

Prediction results for various identified feature sets using the FEAST toolbox, a selection of standard feature-selection methods from machine learning, and our proposed algorithm are shown in Fig. 5. The algorithms we compared our approach against, do not provide a stopping criterion, but only rank features by their importance. Hence, for each algorithm we predicted the fitness using various sets $\mathbf{S}$ of the highest-ranked variables to allow for comparison to our solution. The plots show the prediction error from various sets of sizes up to 10, i.e., $|\mathbf{S}| = 1, \ldots, 10$. In many cases, the feature sets from standard machine learning approaches did not provide accurate predictions. Only for run C and D with $N_{feat} = 18$, run A with $N_{feat} = 45$ and for runs B and C with the largest feature set, the predictions of one standard method allowed for rather accurate predictions compared to selected feature sets of the same size. The features sets selected with the MRMR method, as one of the best performing methods from the FEAST toolbox, performed quite well, but only when a small number of features was selected and the relative performance dropped for larger sets of selected features. The proposed method based on CMI feature selection performed well for all studied situations as it consistently gave a good trade-off solution with respect to feature set size and prediction accuracy. In 14 out of the 16 considered runs and number of features, our algorithms selected the best feature set among all feature sets of the same or smaller size. In 6 of these cases, the selected feature set performed best across all feature sets of any size. The other methods did not provide feature sets with such consistently good prediction performance, as can be seen, for example, for the LARS (blue crosses) and MRMR (red crosses) method in Fig. 5, which performed well for some configurations, but did not return good results consistently.

Generally, we observed that many different feature sets led to similar prediction performance, especially for the largest feature set with $N_{feat} = 120$, which supports our previous analysis that this parametrization lead to highly redundant and correlated features. Nevertheless, the proposed CMI-based feature selection algorithm still managed to identify meaningful feature sets which were not too large and which allowed for good prediction performance.

IV. Discussion

We applied a recently introduced information-theoretic approach to feature selection [10] in sensitivity analysis for optimization data. A strong conceptual and practical advantage of the proposed feature selection approach is its ability to account for interactions between variables when selecting features, such that the selection of redundant features is avoided while features that contribute information in a synergistic fashion together with other features are included. A further significant advantage of the used algorithm for the present application is the ability to automatically determine the number of relevant features by means of statistical testing, whereas for most established methods the number of features has to be fixed in advance. Furthermore, we used the recently introduced partial information decomposition (PID) framework [17] to identify feature interactions.

We successfully applied the approach to four realistic aerodynamic optimization runs, where we showed that the feature sets identified by the proposed algorithm always provided a good trade-off solution with respect to feature set size and prediction performance. We showed that in most of the cases (14 out of 16) the selected feature set could be used to predict the optimization's objective function with smaller error than using feature sets of the same size or smaller identified through existing approaches.

Central to the proposed approach is its ability to identify feature sets while accounting for interactions between features and to identify synergistic interactions. This property is especially desirable in application domains where optimization parameters are expected to show interactive effects on the target function. Such an analysis was previously not possible using the MI or its extensions, for example, the interaction information [34], which was proposed for the analysis of interactions in design data in earlier studies (e.g., [1], [2]). However, it was shown that these measures are not able to disentangle redundant and synergistic contributions and that such a contribution required the axiomatic extension of classical information theory as was done in the PID framework [17] (see also [18]). Accordingly, the development of information-theoretic filtering methods accounting for interactions has not advanced in recent years such that the methods employed here, which often assume variable independence, are still a common approach (e.g., MRMR [35], [36]). We believe that

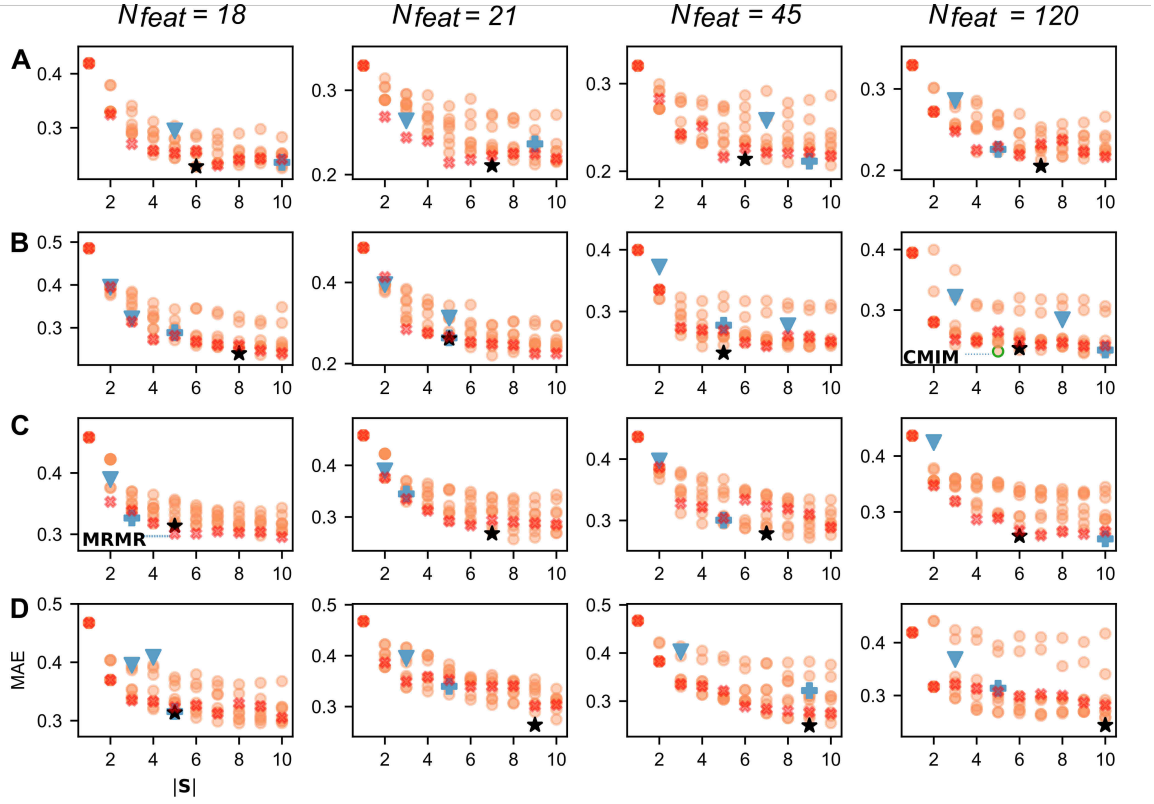

Fig. 5. Validation of feature set through k-nearest-neighbor-prediction of optimization outcome from selected feature sets (mean absolute error, MEA, y-axis). The x-axis indicates the size of the feature set, $|\mathbf{S}|$ (see main text). Each row shows results for one optimization run A-D. Each column shows one feature set with $N_{feat} = 18$, 21, 45, and 120 features. Orange markers indicate prediction results for FEAST feature sets of different size, blue markers indicate prediction results from machine learning approaches. Blue crosses are the feature sets selected with LARS while red crosses are those selected by MRMR. The black star (★) indicates results from our algorithm. Annotations indicate the two cases where another method gave better prediction results than our method for a smaller or equally large feature set.

this stagnation is partially due to the inherent lack in classical information theory to describe multivariate information contributions that has only become available with the introduction of PID [10], [17]. Hence, PID enables the information-theoretic quantification of interactions in design applications as defined in [6]: "a design interaction is defined as a unique dependency between design and objective parameters from which all dependencies of lower ordinality are removed".

The algorithm used for feature selection employs statistical testing to handle the bias in information-theoretic estimates. Statistical testing furthermore provides an automatic stopping criterion as it can reveal that an estimate is not significantly different from an estimate from data with no relationship. Using statistical testing in feature selection has been proposed, for example, by [37]. However, the approach used here is the first to rigorously control the family-wise error rate when testing repeatedly during iterative feature selection [9].

The used algorithm accounts for redundant and synergistic contributions during the identification of relevant features by conditioning on the set of all already selected features. A limitation is here that due to the iterative inclusion, variables that provide purely synergistic information can not be detected. To handle this latter scenario, one may start feature selection with a non-empty set, e.g., some random subset or a subset informed by prior information. Alternatively, one may include variable tuples instead of individual variables [38].

A further limiting factor is the number of features that the algorithm is able to select given a certain amount of data. If the selected feature set becomes too large, CMI-estimation suffers from the curse of dimensionality such that the CMI can no longer be estimated reliably from the available. As a result, the estimate fails to reach statistical significance and the algorithm terminates. However, in sensitivity analysis it is typically the goal to identify the set of *most relevant* features that can still be meaningfully interpreted by a human. As shown here, the algorithm was able to identify up to 10 informative variables from less than 2000 highly biased samples.

Regarding the engineering task of identifying the most influential regions of the shape design the proposed approach gave satisfactory results, as features located at known highly influential region were successfully identified. Also, the high degree of redundancy and correlations in the features sets, which is a natural consequence of the smoothness of the shape deformations, is handled well by the approach. Future work may focus on a visualization and interpretation of the results to provide a more intuitive picture to the engineer who is

potentially not vell-versed in information theory.

We conclude that the proposed algorithm [8]–[10], together with the recently introduced PID framework [17]–[19] and suitable estimators [5], [23], provides a valuable tool for the assessment of optimization outcomes in practical applications. In particular, the interaction-aware feature selection together with the estimation of synergistic effects allows to identify interactions between optimization parameters that was previously not possible using information-theoretic methods. Thus, the novel extension to information-theoretic analysis presented here provides powerful tools for quantifying relationships in a wide area of application domains that are concerned with the analysis of data from non-linear systems.

REFERENCES

[1] L. Graening, M. Olhofer, and B. Sendhoff, "Interaction detection in aerodynamic design data," in *International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2009). Lecture Notes in Computer Science.* Springer, 2009, vol. 5788, pp. 160–167.

[2] M. Rath and L. Graening, "Modeling design and flow feature interactions for automotive synthesis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6936 LNCS, pp. 262–270, 2011.

[3] J. Kmec and S. Schmitt, "Exploring the fitness landscape of a realistic turbofan rotor blade optimization," in *6th International Conference on Engineering Optimisation (EngOpt 2018)*, 2018.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[5] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, p. 16, 2004.

[6] L. Graening and B. Sendhoff, "Shape mining: A holistic data mining approach for engineering design," *Advanced Engineering Informatics*, vol. 28, pp. 166–185, 2014.

[7] L. Graening, S. Menzel, T. Ramsay, and B. Sendhoff, "Application of sensitivity analysis for an improved representation in evolutionary design optimization," *IEEE International Conference on Genetic and Evolutionary Computing*, pp. 1–4, 2012.

[8] P. Wollstadt, J. T. Lizier, R. Vicente, C. Finn, M. Martínez-Zarzuela, P. A. M. Mediano, L. Novelli, and M. Wibral, "IDTxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks," *Journal of Open Source Software*, vol. 4, p. 1081, 2019. [Online]. Available: https://github.com/pwollstadt/IDTxl

[9] L. Novelli, P. Wollstadt, P. A. M. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *Network Neuroscience*, vol. 3, pp. 827–847, 2019.

[10] P. Wollstadt, S. Schmitt, and M. Wibral, "A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition," *arXiv preprint, arXiv:2105.04187 [cs.IT]*, 2021.

[11] R. M. Hicks and P. A. Henne, "Wing design by numerical optimization," *Journal of Aircraft*, vol. 15, pp. 407–412, 1978.

[12] N. Hansen, "The CMA evolution strategy: a comparing review," in *Towards a new evolutionary computation. Studies in Fuzziness and Soft Computing, vol 192.* Berlin: Springer, 2006, pp. 75–102.

[13] E. A. Baskharone, *Principles of Turbomachinery in Air-Breathing Engines*. Cambridge, UK: Cambridge University Press, 2014.

[14] H. Rusche and S. Schmitt, "Stability improvements of pressure-based compressible solver and validation for industrial turbo machinery applications," in *4th Annual OpenFOAM User Conference*, 2016.

[15] L. Graening, S. Menzel, M. Hasenjäger, T. Bihrer, M. Olhofer, and B. Sendhoff, "Knowledge extraction from aerodynamic design data and its application to 3D turbine blade geometries," *Journal of Mathematical Modelling and Algorithms*, vol. 7, pp. 329–350, 2008.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[17] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv Preprint arXiv:1004.2515 [cs.IT]*, 2010.

[18] A. J. Gutknecht, M. Wibral, and A. Makkeh, "Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic," *Proceedings of the Royal Society A*, vol. 477, p. 20210110, 2021.

[19] A. Makkeh, A. J. Gutknecht, and M. Wibral, "Introducing a differentiable measure of pointwise shared information," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 103, p. 032149, 2021.

[20] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[21] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, p. 026209, 2007.

[22] G. Doquire and M. Verleysen, "A comparison of multivariate mutual information estimators for feature selection," in *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, 2012, pp. 176–185.

[23] A. Makkeh, D. O. Theis, and R. Vicente, "BROJA-2PID: A robust estimator for bivariate partial information decomposition," *Entropy*, vol. 20, p. 271, 2018.

[24] T. Sonoda, R. Schnell, T. Arima, G. Endicott, and E. Nicke, "A study of a modern transonic fan rotor in a low reynolds number regime for a small turbofan engine," in *Turbo Expo: Power for Land, Sea, and Air. Volume 6A: Turbomachinery*, vol. 55225, 2013, p. V06AT35A032.

[25] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, "Conditional likelihood maximisation: A unifying framework for mutual information feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[26] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems (NIPS '99)*, vol. 12, 1999, pp. 687–693.

[27] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 212–217.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1226–1238, 2005.

[29] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[30] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Workshops on Applications of Evolutionary Computation*. Springer, 2006, pp. 91–102.

[31] D. Lin and X. Tang, "Conditional infomax learning: an integrated framework for feature extraction and fusion," in *European Conference on Computer Vision (ECCV 2006)*. Springer, 2006, pp. 68–82.

[32] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissertation, Univerza v Ljubljani, 2005.

[33] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*. Morgan Kaufmann, 1992, pp. 249–256.

[34] W. J. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, pp. 97–116, 1954.

[35] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, "Streaming feature selection algorithms for big data: A survey," *Applied Computing and Informatics*, 2020.

[36] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[37] A. Tsimpiris, I. Vlachos, and D. Kugiumtzis, "Nearest neighbor estimate of conditional mutual information in feature selection," *Expert Systems with Applications*, vol. 39, pp. 12 697–12 708, 2012.

[38] J. T. Lizier and M. Rubinov, "Multivariate construction of effective computational networks from observational data," *Preprint no.: 25/2012, Max Planck Institute for Mathematics in the Sciences*, 2012, available from: https://www.mis.mpg.de/publications/preprints/2012/prepr2012-25.html (accessed: 2020-12-15).