

Effective Decision Making and Data Visualization Using Partitive Clustering and Principal Component Analysis (PCA) for High Dimensional Pareto Frontier Data

Saket Kansara
ESTECO Software India Pvt. Ltd.

Sumeet Parashar and Zhendan Xue
ESTECO North America Inc

ABSTRACT

Decision making in engineering design is complicated, especially when dealing with high-dimensional data. Modern software tools are able to produce a large amount of data while performing optimization studies. A typical optimization problem with many objectives may produce 100s or even 1000s of Pareto Optimal solutions. It is a challenge to analyze this data and make a decision about which design/s to choose for further testing or as a final design.

To tackle the problem, two data analysis techniques are used in this paper. Partitive Clustering (PC) is used to locate groups of similar designs in the dataset while Principal Component Analysis (PCA) is used to reduce the dimensionality of the data and visualize it in two and three dimensions. Although these techniques can be used independently, when used together, they prove to be a tremendous help in decision making. This paper underlines the benefit of using these two methods together.

Two multi-objective engineering design optimization problems are chosen to demonstrate the use of PC and PCA together. It is assumed that optimization has been performed and Pareto data is available. The task is to choose a handful of representative designs from this set for further analysis. The aforementioned technique of using Partitive Clustering and Principal Component Analysis together is applied to dataset for both problems. It will be shown how PC+PCA together can easily and effectively reduce the decision set while preserving the overall trade-off. Fewer design choices and fewer decision attributes can be visualized and communicated more easily and can certainly help the decision maker to pick a reasonable final solution.

CITATION: Kansara, S., Parashar, S., and Xue, Z., "Effective Decision Making and Data Visualization using Partitive Clustering and Principal Component Analysis (PCA) for High Dimensional Pareto Frontier Data," *SAE Int. J. Mater. Manf.* 8(2):2015, doi:10.4271/2015-01-0460.

INTRODUCTION

Increased sophistication of engineering simulation tools, availability of computing power and demand for optimization is steering engineers towards multi-objective and multi-disciplinary formulation. It is becoming common to see problem setup with hundreds of input parameters, many objective functions and constraints. While modern optimization algorithms and software tools are able to handle large scale problems and generate even larger datasets, it is a challenge to analyze such large data sets and make effective decisions. Decision making is especially challenging when one has multiple objectives of conflicting nature. Pareto frontier solutions for two or three objective space are easy to visualize and comprehend, however, as the number of objectives increase, it becomes more and more difficult to understand and analyze trade-offs from the results coming out of optimization. Often, optimization problems with three or more objectives can result in hundreds or thousands of Pareto points.

Despite all the advancements in software tools and decision making methods, it can be a daunting task to understand trade-offs and narrow down choices for let's say 10 objectives from a Pareto set of 100s to 1000s of data points.

Multi-Variate data Analysis (MVA) methods such as Self Organizing Maps (SOM) [1], Principal Component Analysis (PCA) [2][3], Data Clustering etc have been used independently in variety of data analysis applications from economics, social sciences etc. However, use of such methods for engineering data analysis and decision making has been very few [2][4][5]. Clustering techniques can help identify groups of similar designs in a dataset and try to compress the size of data set, while 'Principal Component Analysis' can help reduce the dimensionality of the data by representing correlated variables and projecting data on two or three principal components.

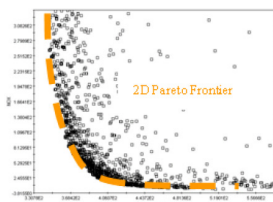
This paper highlights a unique combination of applying Partitive Clustering and PCA together to a high dimensional Pareto frontier data set for effective decision making. This method has been successfully applied to two different case studies. The first consists of a multi-objective Pareto frontier data set of over 150 design samples for 5 objectives. The second case contains a more intricate Pareto frontier data set of over 6000 designs in a 12 objective design space. The paper explains how the clustering method helps reduce the decision set to representative finite choices to choose from while still covering the Pareto frontier spread. The paper also showcases the subsequent use of PCA to reduce the dimensionality to two or three representative dimensions while still representing the relative positioning of original variables. The study also illustrates effective visual charts used to visualize, interact and make decisions based on this Clustering + PCA approach.

DIFFICULTIES IN DECISION MAKING AND VISUALIZATION USING HIGH DIMENSIONAL DATA

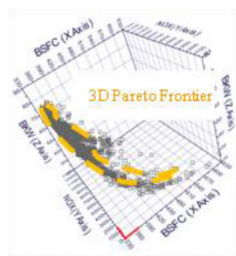
Figure 1 below contains results from different optimization problems with number of objective functions ranging from one to three. If the optimization problem contains a single objective, it yields a single optimum design. As can be seen from Figure 1a, it is quite straightforward to choose the optimum design in this case. In case of a two objective optimization problem, a two dimensional Pareto frontier of designs is generated. As shown in Figure 1b, it becomes little more involved effort to choose a final optimum design since individual preferences for the objectives need to be accounted. However for two objectives, it is still manageable to see and understand the trade-offs. If there are three objectives, it leads to a three dimensional 'surface' of Pareto frontier designs as shown in Figure 1c and in this case it becomes even more difficult to visualize, understand and choose the final optimal design. If four objectives are present, it produces a four dimensional hyper-surface of Pareto frontier designs and so on. Hence, it becomes more and more difficult to visualize and analyze Pareto frontier designs as the number of objectives increases.



a. Single Objective Optimization



b. Two-Objective Optimization



c. Three Objective Optimization

Figure 1. Results from One, Two and Three Objective Optimization

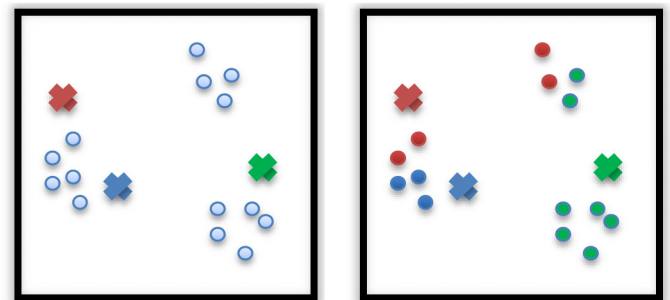
Hence, in modern engineering design optimization problems, where one has many objectives, visualizing such high dimensional data and making decisions using this data can become quite a challenging task.

The rest of the paper is structured as follows. The next section provides a brief introduction to Partitive Clustering method and PCA methods. Afterwards the approach of using Clustering and PCA together for effective data analysis is explained using two example cases.

INTRODUCTION TO PARTITIVE CLUSTERING

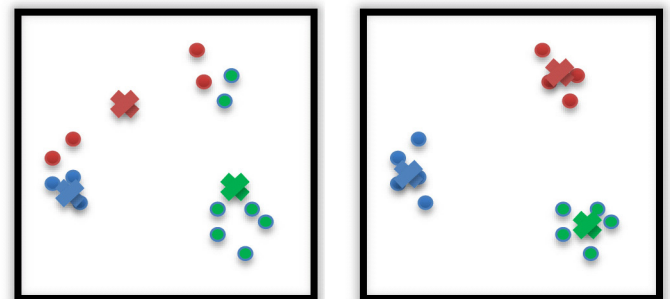
The aim of clustering is to find structures in data. A cluster is a subgroup of data which are "similar" among them and "dissimilar" with respect to data belonging to other clusters. Clustering methods can be divided into two basic types: Hierarchical and Partitive clustering.

Partitive Clustering is one the most popular clustering algorithms. It divides the data in a number of predefined partitions in such a way that the Euclidean distance between the instances and their respective cluster centroid is less than the distances to any other centroid. In this type of clustering analysis the user must specify in advance the desired number of k clusters. After identifying the initial centroids, the algorithm iterates by grouping samples based on the nearest Euclidean distance from the cluster centroids, seeking to minimize the variation within the clusters. The process is performed until cluster centroids do not shift for more than a given threshold or until the maximum number of iterations is reached.



Step 1: Initial Guess is 'Three'. Start with Random Centroids

Step 2: Assign each instance to closest centroid



Step 3: Move centroids to center of assigned points

Step 4: Iterate until convergence is reached

Figure 2. Partitive Clustering - explained step by step

Figure 2 explains the Partitive Clustering algorithm in detail. The 'Blue bubbles' in different subfigures represent different design points. To initialize the clustering algorithm, in step 1, we start with an initial guess of 'three' clusters and generate three random centroids. In step 2, the distance of each design is measured from each centroid and designs closest to a centroid are assigned to that particular centroid. In step 3, the centroid is moved from its original location to a new location calculated based on the newly formed cluster of designs. This process is repeated until convergence is reached, as shown in step 4.

INTRODUCTION TO PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a mathematical procedure using an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, it accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

PCA is used for representing high-dimensional data in a lower-dimensional form without losing too much information and eliminating redundancies. It is also used for identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences since patterns may be hard to find in high-dimensional data.

Fig. 3 shows an academic example for two variables y_1 and y_2 which are positively correlated. However, notice that most of the variation in the data occurs in a direction along the main diagonal, and only a small amount of variance occurs in a direction orthogonal to this one. These directions are known as principal components and are shown by the arrows. PCA is a change of coordinates that rotates the basis in Figure 3(a) onto the principal components in Figure 3(b). The advantage of doing so is that correlation is removed in this new basis. It thus becomes possible to neglect latent variables that do not contribute significantly to the variance for only a minor loss of information. Ignoring v_2 , for example, the system can now be described with good approximation using only one dimension v_1 , as shown in Figure 3(c). [2]

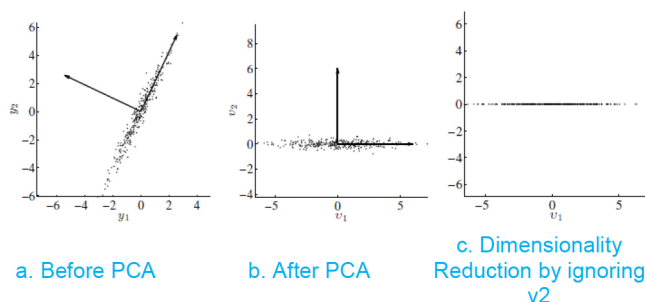


Figure 3. Illustration of Concept behind Principal Component Analysis

In order to perform dimensionality reduction, PCA finds eigenvectors of the dataset. The eigenvalues corresponding to each eigenvector represent the amount of variance explained by that component of PCA. The bigger the eigenvalue, the bigger is the variance captured in that principal component. Figure 4 below shows an eigenvalue plot resulting from a PCA study. λ_1 in the figure represents eigenvalue corresponding to principal component 1, λ_2 represents eigenvalue for principal component 2 and so on. The dataset originally contains ten variables, hence there are ten principal components and hence ten eigenvalues in the plot. The height of each λ bar represents the eigenvalue and hence the variation explained in that direction. For example, λ_1 value is around 0.3 and hence represents around 30% of variation in data. Similarly, λ_2 value is about 0.2, meaning principal component 2 represents around 20% of variation in data and so on. The Blue curve in the chart represents cumulative effect explained by all principal components combined together. As can be seen, λ_1 to λ_6 together explain around 100% variation in the data. This means that a dataset which was originally represented in ten dimensions can now be accurately represented in six dimensions.

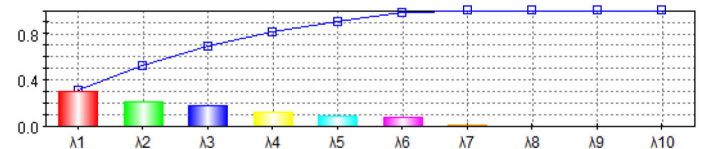


Figure 4. Eigenvalue plot showing all Principal Components explaining variation in data and scope for dimensionality reduction

Figure 5 shows a scatter plot of the point cloud in principal component space with first principal component on X axis and second principal component on Y axis. The scatter chart is a representation of ten dimensional data in two dimensions. Designs close to each other in this chart can be termed as similar designs while designs far away from each other can be viewed as dissimilar designs. Hence, if one has to choose a few representative samples out of this data, they can select designs which are well spread throughout the scatter chart. The green arrows in the chart represent how the original variables are aligned and contributing to the principal components. The length of the arrows represent the magnitude of contribution while the direction of the arrows represent how much a variable contributes to that particular principal component.

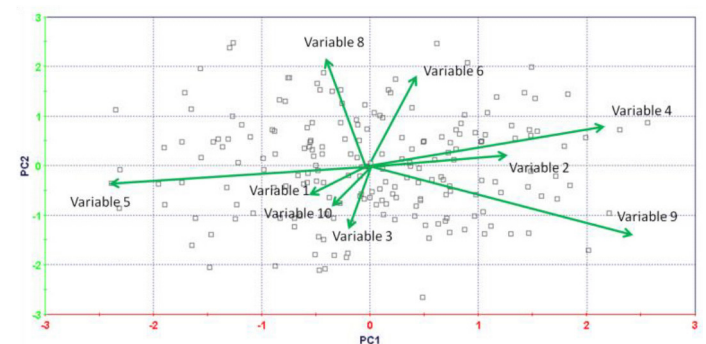


Figure 5. Illustration of Concept of 2D scatter chart based on Principal Component Analysis