44th AIAA Aerospace Sciences Meeting and Exhibit
9 - 12 January 2006, Reno, Nevada

AIAA 2006-727

An Intuitive Approach to Visualize Multidimensional Data and Relationships

G Agrawal* and C. L. Bloebaum†
Department of Mechanical and Aerospace Engineering
University at Buffalo, Buffalo, NY, 14260

Multidimensional relationships exist in almost any discipline. There are numerous visualization methods that enable us to 'see' multidimensional relationships in an easy and intuitive manner for two or three variables, using 2-d and 3-d representations. It is substantially more difficult, and sometimes even impossible, to visualize more than three dimensions – at least easily or in a way that is intuitive to the user. In this paper, a visualization methodology is presented in which multidimensional relationships can be viewed in an intuitive and straightforward manner. Based on this visualization, it is possible to quickly identify regions of interest for functional relationships, optimization applications, or for high dimensional datasets, regardless of the complexity of space or data. The technology provides a new approach to visualize more than three dimensions in a way that looks remarkably similar to traditional 2-d and 3-d representations. The method uses a new technique of 'lossless' dimension blending, termed the Hyper-Space Diagonal Counting (HSDC). The methodology developed here provides a unique way to visually represent the relationships for n-dimensional problems. What is described here represents a totally new methodology that has the potential to greatly impact numerous industries, as well as the educational enterprise.

I. Introduction

INFORMATION from complex and large datasets generated by simulations, experiments, or observations, can be better and quickly understood if the data is presented in an image format instead of just textual. When it comes to understanding the relationships in the data, throughout history, scientists, engineers, and many others have used simple or complex graphs to represent their data visually. While it is easy to understand relationships using two-dimensional graphs, and sometimes even three-dimensional graphs, it is presently very difficult or sometimes even impossible to easily visualize more than three dimensions in an intuitive manner.

With the availability of incredibly powerful computational platforms for relatively low cost, we are seeing an ever-increasing generation of pure data. Whether for engineering analysis, market trend exploration, entertainment simulation, or any other profession, the data exists. It then becomes a tremendous challenge to synthesize and process the data appropriately so as to derive use.

Coupled with the explosion of larger data sets, is the increasing prevalence of multidimensional data. Computational platform availability, along with more sophisticated computational algorithms and incredible processing power, has resulted in a pervasiveness of data sets with hundreds or thousands of dimensions[1]. Without going into detail as to the specific representation, it is obvious that the visualization of Fig. 1 represents a great deal of data, thereby yielding a wealth of information.

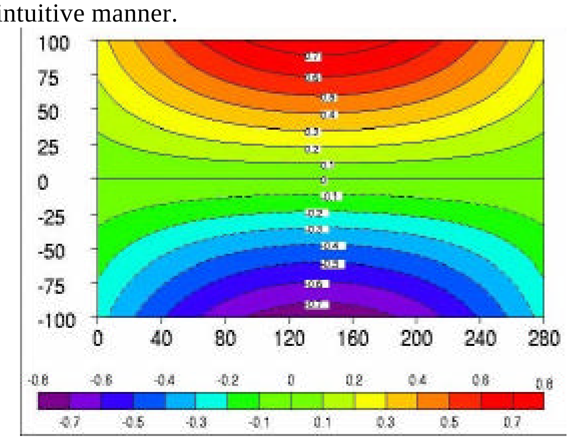


Figure 1. Simple Visualization

Visualization of multidimensional data and relationships represents a particularly unique challenge to the scientific community. Two- and three-dimensional visualization is easy. Using sound, color, and/or motion might

* Research Assistant, MAE Department, University at Buffalo, Student Member AIAA
† Professor for Competitive Product and Process Design, MAE Department, University at Buffalo, Associate Fellow AIAA

enable us to represent five dimensions in a still somewhat intuitive way. Going beyond this, however, while maintaining an intuitive representation of the multidimensional data, is extremely difficult.

Numerous methods and applications have been developed to meaningfully translate large amounts of multidimensional data into intuitive visual representations. The type of the multidimensional data being investigated for a specific application ultimately dictates the mechanism by which information and subsequent knowledge might be gained. For instance, we might wish to explore a multidimensional space in which the dimensionality is dictated by the number of design variables (i.e. independent variables) that impact a defined objective function and set of constraints in an engineering optimization application. Alternatively, we might wish to explore an existing dataset to identify relationships and trends in data. We might say that in one case we are interested in exploring the space defined by the variables and their relation to the metric function, while in the other case we are interested in exploring an existing dataset to identify potential meaningful relationships amongst the data.

This paper describes a new methodology, termed the Hyper-Space Diagonal Counting (HSDC) to enable visualization of a multidimensional data and underlying relationships. The key challenges reside in how to best capture all dimensional relationships and then how to present them in a visual format that will be sufficiently meaningful and intuitive. For visualizing the multidimensional relationships, there exist many different methods, which provide necessary means to visually represent the large amount of information existing in the data. However, with most of these methods, the visual representation results in a loss of meaning, a loss of the concept of neighborhood, and the subsequent loss of ability to understand the representation in an intuitive way. To identify specific issues related to the existing state-of-the-art multidimensional multivariate visualization (MDMV) techniques, the previous work done in this area must be reviewed.

II. Background

The term dimensionality can be somewhat confusing, given its use in a variety of fields to mean different things. Multidimensional objects, for instance, are spatial objects, for which a desire exists to understand their geometry, while multidimensional data suggests relationships amongst numerous variables[2]. Terms such as hyper-dimensional, multidimensional and high-dimensional, have been used to mean essentially the same thing[3]. In this paper, the term multidimensional refers to two or more dimensions.

The goal of multidimensional visualization is to meaningfully translate large amounts of multidimensional data into intuitive visual representations[4]. Numerous methods and applications have been developed to achieve this goal, subject to hardware and software limitations of the 2-d or 3-d visualization space. All multidimensional visualization techniques attempt to transform a multidimensional problem or dataset so that it can be mapped to a 2-d or 3-d visual space through the use of dimension reduction.

A. Dimension Reduction

A standard approach that is used in multidimensional visualization is to first process the multidimensional data in such a way so as to reduce the dimensionality, while still preserving the integrity of its meaning. Fig. 2 shows this implementation.

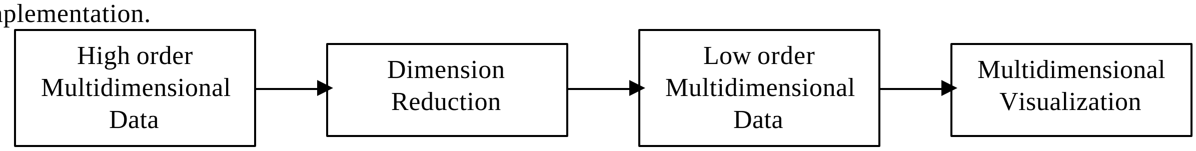


Figure 2. Procedure for implementing dimension reduction

This is a particularly important step for visualization, as the associated complexity of the problem or dataset is significantly reduced. One of the most comprehensive resources for discussion of dimension reduction techniques can be found in Carreira-Perpinan[5]. As pointed out in this paper, dimension reduction techniques assume that reduction is possible for two reasons: 1) Some of the variables are actually irrelevant in that they have variations smaller than measurement noise, and 2) Some variables can be correlated with each other.

Dimension reduction is accomplished in a variety of ways in different fields, including projection techniques[6, 7, 8], data compression[9], feature extraction[10], and regression and smoothing techniques[11, 12], among others. One of the most widely used dimension reduction techniques is Principal Component Analysis (PCA)[13, 14, 15], wherein some number of principal components is found through a symmetric regression approach. Projection techniques are computationally expensive and are most suitable only for linear structures, which represents a significant drawback. These techniques all have various drawbacks that ultimately limit their applicability to a wide range of problems. Further, many of these result in losses of dimensions in the new representation. This is problematic for a variety of reasons, the most important being that one never really knows what secondary and tertiary effects might be lost through the elimination of a dimension considered potentially unimportant. In the HSDC method developed in this work, all dimensions are represented and preserved through blending, rather than the traditional dimension reduction approaches. The next section briefly reviews existing techniques for multidimensional visualization.

B. Multidimensional Visualization Techniques

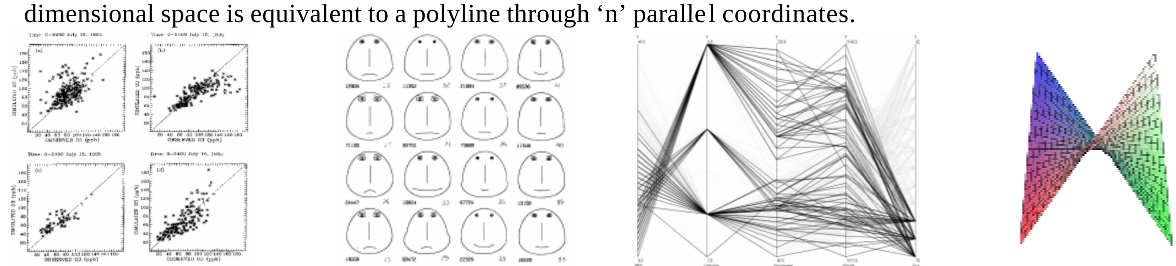
Most multidimensional visualization techniques that depend on dimension reduction in some way result in a loss of meaning, and a loss of ability to understand the representation in an intuitive way. Multidimensional visualization techniques can be categorized in a number of different ways. Possible criteria for such a categorization include the goal of visualization, the type and/or dimensionality of the data, the use of color, the use of animation, and the dimensionality of the visualization technique, amongst many other possibilities. Two broad categories[4] summarized below provide an overview of the best-known multidimensional visualization techniques.

1) Techniques designed for fixed number of variables

In this class of methods, color is generally used as a fourth dimension and time is used as an animation parameter to represent a fifth dimension. This includes fitted curves, reference grids, and banking.

2) Techniques designed for any number of variables

Here, symbols, and matrices of views are often used as a key to the representation. One of the best-known methods is the scatter plot matrix, where n dimensions are projected onto n*(n-1) scatter plots, in which each pair of dimensions has two scatterplots showing their relation[4]. HyperSlice[16] and HyperBox[17] can be considered extensions of scatterplot matrices, in which color and interactivity provide for greater insight into the problem. Another is Chernoff Faces[18] and, more broadly, the use of glyphs[19, 20], to represent characteristics of relationships and the space in question. Hierarchical axis[21, 22], Dimension Stacking[23], and World within Worlds[24, 25] all use the concept of a hierarchical representation in some way for the dimensional relationships. Parallel Coordinates[26, 27, 28] has become an extensively used approach in a variety of fields and applications. In this approach, a point in n-dimensional space is equivalent to a polyline through 'n' parallel coordinates.



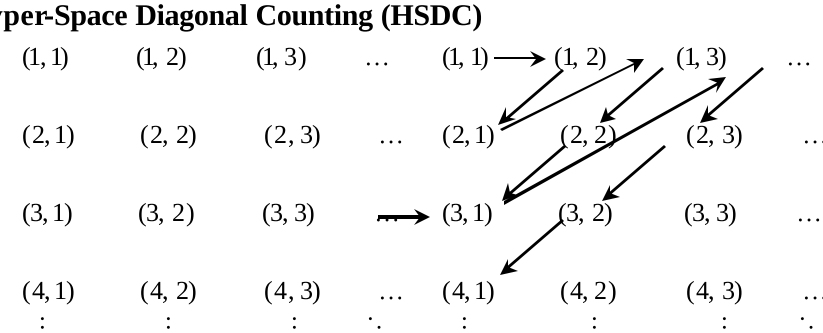
Scatterplot Matrix[4] **Chernoff Faces[18]** **Parallel Coordinates[26]** **Animated Mesh[4]**

Figure 3. Representative visualizations associated with key MDMV techniques

These represent only a small number of methods that have been proposed for multidimensional visualization, but are probably the most widely recognized. Each of the techniques listed under the two broad categories above have advantages and limitations. Some of them are somewhat difficult to understand, some of them are computationally expensive, and some are just awkward. A few of the more widely used techniques are shown in Fig. 3. This paper will attempt to demonstrate the great potentials of visualization based on the Hyper-Space Diagonal Counting (HSDC) method developed in this work.

III. Hyper-Space Diagonal Counting (HSDC)

The concept of counting originated with the famous German mathematician from 19th century, Georg Cantor, who recognized that for every point of a surface, there is a corresponding point of the line and, conversely, for every point of the line there is a corresponding point of the surface. The result of this observation is that there is a one to one correspondence of points on the interval [0, 1] and points in an n dimensional space. It is this correspondence which provides the basis for the development of the Hyper-Space Diagonal Counting (HSDC) method. Cantor's discoveries in set theory rest upon a very simple idea – that even



Array of points on a surface Path through all the points

Figure 4. Graphic proof of Cantor's theory

though you may not be able to count something, you can equate it with something else that has the same cardinality (i.e. the same number of elements). Two infinite sets can be shown to have the same cardinality by finding a one-to-one mapping between the elements of each set. The very famous proof for Cantor's theory of mapping points from a 2-dimensional space on a line can be represented visually in Fig. 4. Consider an array that includes points from a 2-d surface in an order and then make a path through all the points in the diagonal way shown. In this way, every single point can be mapped to one (and only one) point on a line. Hence, for every point on a 2-dimensional surface there is a corresponding point on a line that is unique.

This concept of counting, which represents a very small portion of Cantor's contributions to the field of mathematics, is the key to understanding the Hyper-Space Diagonal Counting (HSDC) method developed here. A path can be created along the points in an n-dimensional space and each point can subsequently be mapped to a single point on a line.

In this work, Cantor's work on counting has been extended to enable mapping points in an n-dimensional space to a line. This mapping of points from an n-dimensional space to a line has been termed the Hyper-Space Diagonal Counting (HSDC) method.

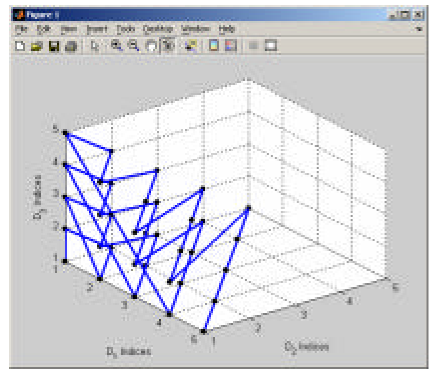


Figure 5. Path for counting in 3-d

Cantor has already provided a proof for 2-d and what remains is a matter of making an array of n^{th} ordered combinations of all the points in an n-d space and then creating a path through all the points to get a sequence. An extension of Fig. 4 to show a mapping for 3-d points is given in Fig. 5. This can be extended to n-dimensions. We can create a path by counting each point in the n-dimensional space and subsequently mapping each of those points to the points on a line, which forms the basis of the HSDC approach. In order to understand how this counting is generalized for n-dimensions, consider Table 1, where the counting has been extended to a *Hyper-Space Diagonal Counting* in four dimensions, by observing vital trends from 2-d and 3-d counting described above. First, the indices of the points counted in 2 and 3 dimensions are laid out in a tabular format. Shades of grey signify the diagonals (in 2-d and 3-d) or the hyper-diagonal in 4-d. Indices of points in the same shade grey band represent the points on the same level (diagonal/hyper-diagonal). Listing the points in the tabular format (shown in Table 1), reveals important trends in the counting identified as follows.

Table 1. Counting Extended to 4-d

2-d Counting		3-d Counting			Extended to 4-d			
D1	**D2**	**D1**	**D2**	**D2**	**D1**	**D2**	**D3**	**D4**
1	1	1	1	1	1	1	1	1
1	2	1	1	2	1	1	1	2
2	1	1	2	1	1	1	2	1
		2	1	1	1	2	1	1
					2	1	1	1
Different shades represent different diagonals								

1) The sum of all the indices at a particular level remains constant, which is one less than the sum of the level and dimensions we are looking at. For instance, in 2-d counting at diagonal 2 (dark grey), all the indices sum to 3 (i.e. 2+2-1=3). Similarly, in case of 3-d counting, at diagonal 2, all indices sum to 4 (i.e. 2+3-1=74), which is clear in the table. This holds true for counting in higher dimensions.
2) If we look at 2-d and 3-d counting, we observe that the number of points at a particular diagonal in 3-d counting is equal to the total number of points up to the same diagonal in 2-d counting. For instance, at the second diagonal in 3-d counting there are 3 points, which is the same as the total number of points up to the second diagonal in 2-d counting. This also holds true for counting in higher dimensions.
3) Finally, the way the counting progresses has a trend in itself.

These observations allow us to easily extend the 2-d and 3-d counting into the Hyper-Space Diagonal Counting, as shown by the counting extended to 4-d in Table 1. The same procedure can be followed for counting beyond 4-d. In fact, this can be extended to any number of dimensions. To automate the counting for n-dimensions, the following terms have been defined. The variable '**n**' is used to denote the number of dimensions, '**l**' is used to represent the diagonal on which a particular point falls (i.e. which diagonal/hyper-diagonal out from the origin), '$\mathbf{E_l}$' is the number of points (i.e. points) on a particular diagonal, '$\mathbf{TE^n_l}$' are the total number of points up to diagonal **l**, '$\mathbf{i^l_n}$' is the index of an point for a particular dimension on a diagonal, '$\mathbf{S_l}$' is the sum of all the indices of a point on a particular diagonal in an n-dimensional space, and '$\mathbf{i^l_{max}}$' is the maximum value of the index for any dimension at a particular diagonal. There are several equations (1 through 6) that have been developed in order to be able to automate the counting in n-dimensional space. These are presented on the next page.

1) The equation for the total number of points at a particular diagonal:

$$E_l^n = \left(\prod_{k=0}^{k=n-2} (l+k) \right) \Big/ (n-1)! \qquad \dots n > 1$$

2) The equation for the total number of points up to a particular diagonal:

$$TE_l^n = \sum_1^l E_l^n$$

3) The equation for the points at a diagonal in terms of points at the previous diagonal:

$$E_{l+1}^n = \left(\frac{l+n-1}{l} \right) E_l^n$$

4) The equation for the total number of points up to a diagonal in terms of points at that diagonal;

$$TE_l^n = \left(\frac{n+l-1}{n} \right) E_l^n$$

5) The equation for the sum of all the indices at a particular diagonal:

$$i_1^l + i_2^l + i_3^l + \dots + i_n^l = S_l \quad = \quad (n+l-1)$$

6) The maximum value of the index for any dimension at a particular diagonal:

$$i_{\max}^l \quad = \quad l$$

At this point, many people might be wondering what any of this has to do with visualization. To clarify and present the proper context, the next sections show how the HSDC can be used for visualization of multidimensional space and exploration of datasets, in order t understand the underlying relationships and reveal hidden patterns.

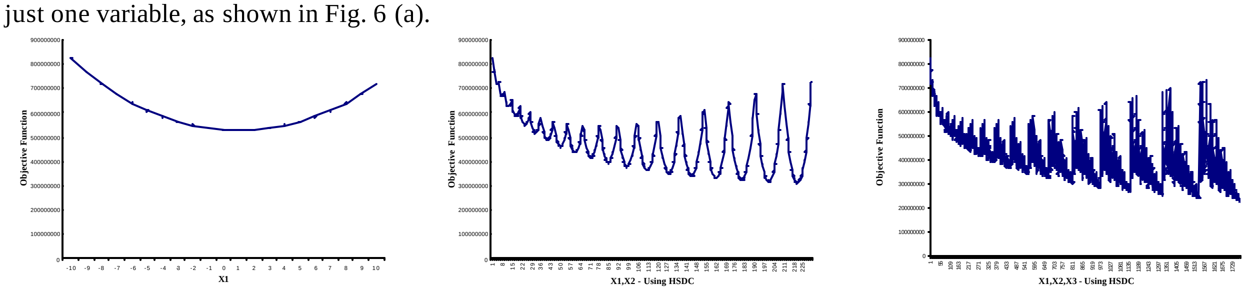
IV. HSDC's Application to Visualize Multidimensional Functional Relationships

Multidimensional functional relationships exist in almost any discipline. An example would be how the width, length, and thickness of a piece of material directly affect the overall volume of an object. This is a functional relationship. There are numerous visualization tools that exist on the market to enable one to 'see' such relationships in an easy and intuitive manner for 2 variables and 3 variables, using 2-d and 3-d representations. It is substantially more difficult, and even impossible in most cases, to visualize more than 3-dimensions – at least easily or in a way that is intuitive for the user. The HSDC can be used to visualize multidimensional functional relations in a way that looks remarkably like that of traditional 2-d and 3-d representations.

Consider the single objective unconstrained minimization problem with six design variables given below.

$$\text{Minimize: } F = 10\sum_{i=1}^{6}(16-i)(X_i - 1)^2 + 10\left[\sum_{i=1}^{6}(16-i)(X_i-1)^2\right]^2 \text{ where, } -10 \le X_i \le 10 .$$

The above problem is a seven dimensional test problem with six of the variables (X's) contributing a dimension each and the function value (F) contributing one dimension. To 'see' the relationship between the function and the variables, one way would be to freeze five variables (i.e. make them constant) and plot the function with respect to just one variable, as shown in Fig. 6 (a).



(a) (b) (c)

Figure 6. One variate visualization of 1-, 2-, & 3-dimensional representations

In part (a) of Fig. 6, we see the result of setting five variables at constant values and plotting the 6th variable in HSDC space (i.e. x-axis represents the index value) versus the objective function value. This however does not show the complete picture of the variation of all the variables with respect to the function value. What is depicted is nothing but a two-dimensional cut of the function and the graphical image depends on the constant values of the frozen variables used to determine the cut. This could be highly misleading in context of an optimization as different values of the frozen variables can give completely different picture of the section being shown. Fig. 6 (b) shows the plot generated by mapping the points in dimensions X_1 and X_2 to the X-axis using the HSDC counting scheme, while freezing the rest four variables. The 'F' value is plotted on the Y-axis. The plot clearly gives a better picture of the functional relationship as two of the variables are varied simultaneously and mapped on to a line using a sequence as discussed previously. It shows the internal bell shaped behavior along with the overall concavity depicted in the previous figure. With an optimization point of view and with such behaviors clearly depicted using this representation, one can easily point out the maximum and minimum values of the function, without any trouble. Fig. 6 (c) shows the plot generated by mapping the points in dimensions X_1, X_2 and X_3 to the X-axis using the HSDC counting scheme, while freezing the remaining three variables. The 'F' value is still plotted on the Y-axis.

This plot gives an even better picture with three of the variables being varied simultaneously and mapped on a line using a sequence of points in three dimensions. We can practically map all the six dimensions to the X-axis and get and overall picture of this 7-dimensional problem using this approach. It would be a complete representation of the function's relationship with all the variables. We can make certain observations from parts (b) and (c) of Fig. 6 such as: 1) trends can be identified in relation to indices; 2) minimum points can be easily identified, even for the n=3 case; and 3) neighborhoods are identifiable.

The variability introduced as a result of counting along the hyper-diagonal, which means there are inherent jumps in function value along a hyper-diagonal, can be ameliorated by a representation as shown in Fig. 7. Here, the red line is a representation of Fig. 6b, through index number 15, while the purple line shows the same data, grouped by level (which corresponds to the x axis). This representation is valuable since the localized variability on each level is smoothed so that the overall trend of the data is more easily understandable. Such a representation is easily extensible to greater dimensions.

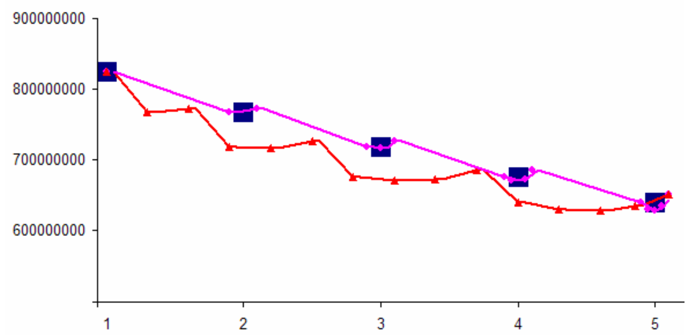


Figure 7. HSDC-based visual representation by level

Alternatively, we can map a few dimensions on the X-axis, a few on the Y-axis and the objective function on the Z-axis. This would result in a hyper-surface representation as shown through a 6-d hyper-surface in Fig. 8 The objective function used for this problem is the equation of a six-dimensional plane $F = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$), where X_i>= 0.0, for i=1,6. Color is used in a traditional temperature scale with red indicating a large value of the function and blue a low value.

Although, in reality this is not a true surface, as we have mapped points from different dimensions to a single axis, it does reflect an image of the functional relationship in the form of a hyper-surface. Such a hyper-surface is much better than having no visual representation at all, or one which is not at all intuitive. This illustrates the powerful potential of applying counting to achieve multidimensional visualization. It is obvious that there are issues such as non-uniformity in mapping of points on the axis, scalability of the approach, etc. but the HSDC approach has great promise for providing a method to obtain an intuitive and meaningful representation of multidimensional multivariate data and relationships.

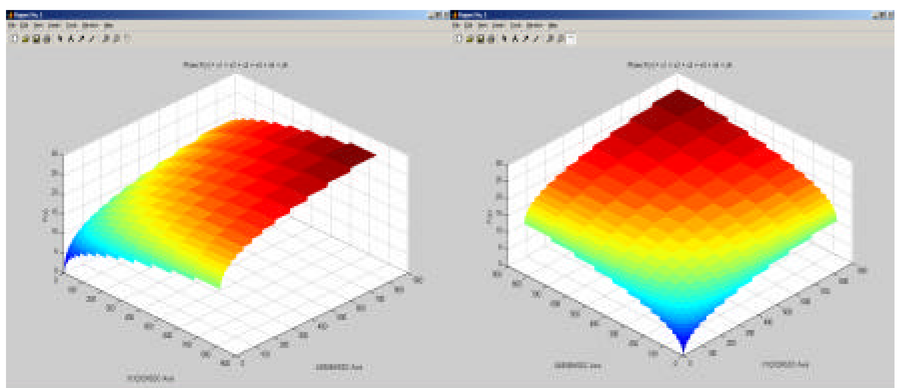


Figure 8. HSDC-based visual representation

The most important point that must be realized is that using the HSDC-based mapping, as we progress from left to right on the mapped axis, there is an overall increment in the values of the indices mapped from multidimensional space. Although the indices themselves (in counted space) increase and decrease at particular diagonals, by virtue of the counting itself, we still have an overall increase from left to right. This is very important in the sense that it is the overall increase of the indices that gives the intuitive understanding to the representation (i.e. the increase from a low to high on the mapped axes).

A. Test Case 1 – Six Variable Product

The objective function used for this problem is the product of all the variables in six dimensions given by: ($F = X_1 \times X_2 \times X_3 \times X_4 \times X_5 \times X_6$), where X_i>= 0.0, for i=1,6. This is a very simple functional relationship for which we know both the maximum value (i.e. all X at infinity) as well as the minimum value (i.e. all X at 0.0). Although we intuitively understand this, we would have no easy way of visualizing such a relationship.

The HSDC approach enables us to group 3 variables per axis and then represent the hyper-surface associated with this problem in indexed space as we see in Fig. 9. As with the introductory problem above, we can now easily see which indices correspond to the lowest value of F and which correspond to the largest value(s) of F. We can easily trace these indices back to the associated X values.

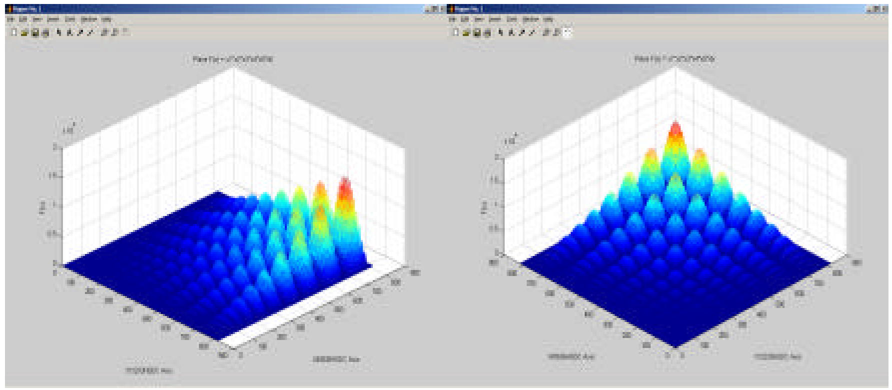


Figure 9. HSDC-based visual representation

The periodic behavior observed in this figure is once again due to the fact hat we count in hyper-diagonals (per level) so that there is inevitably a discontinuous jump in function value as we increase the index number by a single value (i.e. observed most easily in Fig. 4). However, since the HSDC approach is a lossless representation, associated with discretized multivariate space, every piece of data is represented.

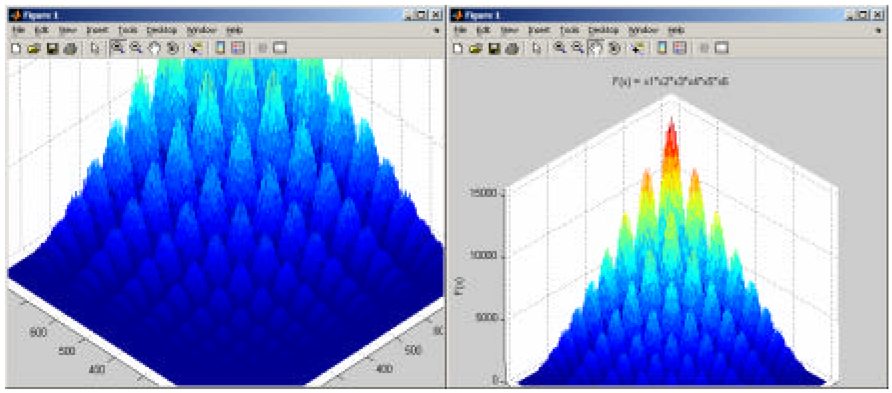


Figure 10. HSDC-based visual representation

One can zoom in on regions of interest and click on any point, which would then result in the associated indices from which the multivariate values can be easily obtained. In Fig. 10, we can see images associated with zooming in on both the minimum function values for this problem, as well as the maximum. It is easy to see that the (0,0) indices correspond to the smallest F value and one can also identify the indices (741,741) as a maximum. The (0,0) indices corresponds, in this case, to all X's with a value of 0.0. The (741,741) indices corresponds to all X's with a value of 5, and a discretization of each variable of 6. Another rotated view of the representation is shown in Fig. 11, wherein it is easier to identify the index of 741 associated with the X4X5X6 HSDC Axis. The function value associated with the maximum point shown here is 15, 625. The HSDC representation, due to its very simplicity, enables interactive selection of indexed points so that the X and

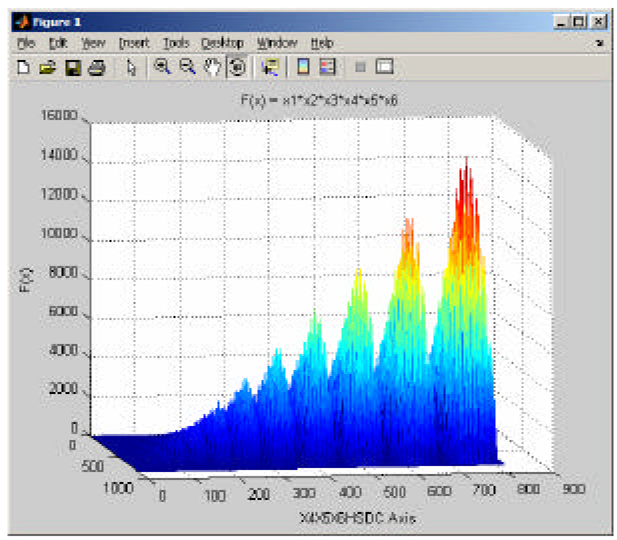


Figure 11. HSDC-based visual representation

B. Test Case 2 – Six Variable Sum of Squares

The objective function used for this problem is the product of all the variables in six dimensions given by: $(F = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 + X_6^2)$, where X_i>= 0.0, for i=1,6. Again, we have a counting of X_1-X_3, and X_4-X_6 on the x and y axes, respectively, with the function value on the z axis.

The same periodic behavior is observed in Fig. 12 for this problem.. Again, we can easily identify 'large' values of the function, as well as 'small' values of the function. We can easily see where the largest value of F is as well as the smallest value. By identifying the indices of interest, we can immediately determine the associated X values for all variables. The image can be rotated and we can 'zoom' in on any region, as with the previous problem. While this is another problem for which we intuitively understand the maximum and minimum, it is illustrative of the ease with which HSDC can be used for such multivariate visualization.

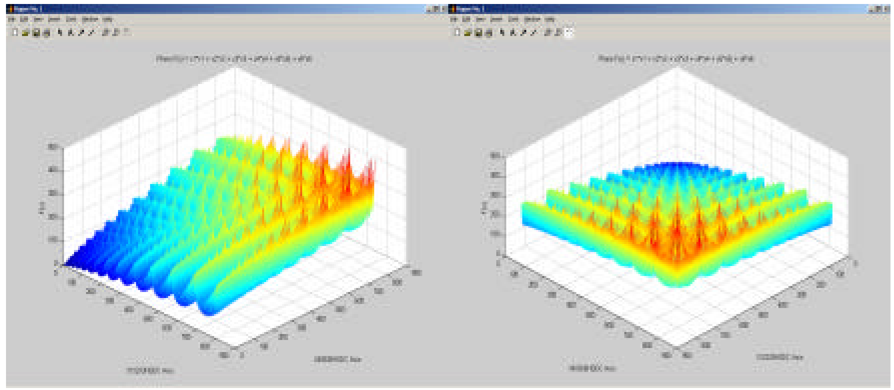


Figure 12. HSDC-based visual representation

C. Test Case 3 – Six Variable Golinski's Speed Reducer Optimization Problem

One application of the HSDC would be for optimization problems. We can 'see' the objective function or constraints in however many dimensions we have associated with the problem. Consider a standard test optimization problem of the Golinski's Speed Reducer. Here, there are 7 design variables, and 25 constraints, with the objective of finding the minimum gearbox volume. The objective functions for this problem is given by:

$$F = 0.7854x_1x_2^2(3.3333x_3^2 + 14.9334x_3 - 43.0934) - 1.5080x_1(x_6^2 + x_7^2) + 7.4770(x_6^3 + x_7^3) + 0.7854(x_4x_6^2 + x_5x_7^2).$$

The full problem statement for this problem can be found in reference [32]. Here, for simplicity, design variable '7' was set a priori at its known optimal value, leaving a 6 design variable problem.

The Hyper-Space Diagonal Counting Method was applied to this problem, yielding the representation for the objective function as shown in Fig. 13. Here, since only the objective function is being represented in the hyperspace visualization, we see that the lowest set of indices corresponds to the smallest objective. However, this representation does not take constraints into consideration. In Fig. 13, indices for each variable vary from 1 to 16, resulting in a relatively smooth hyper-surface as shown. In Fig. 14, indices vary from 1 to 9, resulting in a much coarser discretization of the space. A discussion of impact of discretization on HSDC representation can be found in reference [31].

By considering only those function values which correspond to feasible points in the original design space, the representation of Fig. 15 is

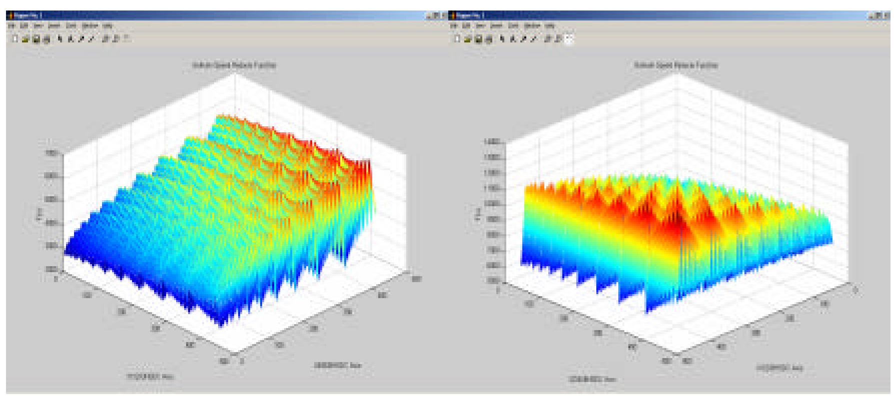


Figure 13. HSDC-based visual representation for speed reducer problem

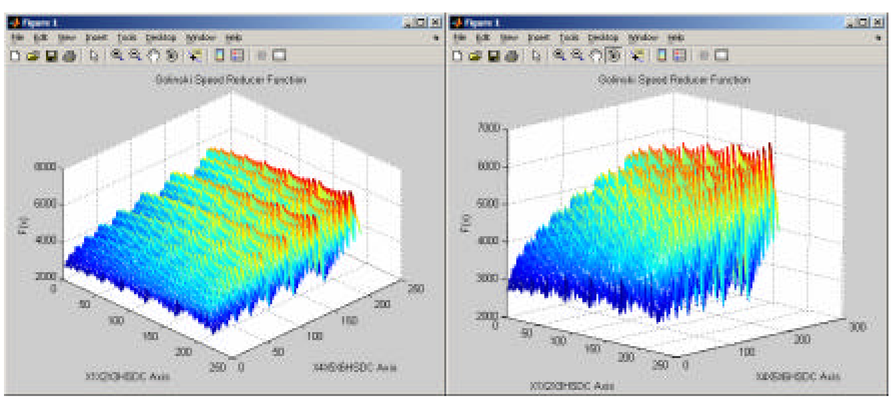


Figure 14. HSDC-based visual representation for speed reducer problem

obtained. A quick rotation of the image yields the smallest feasible value of F of 3089, which corresponds to indices (1,69). From this information, the associated design variable values are very easily obtained. Here, then, we have been able to easily visualize a constrained optimization problem with 6 design variables and quickly identify the minimum feasible function value.

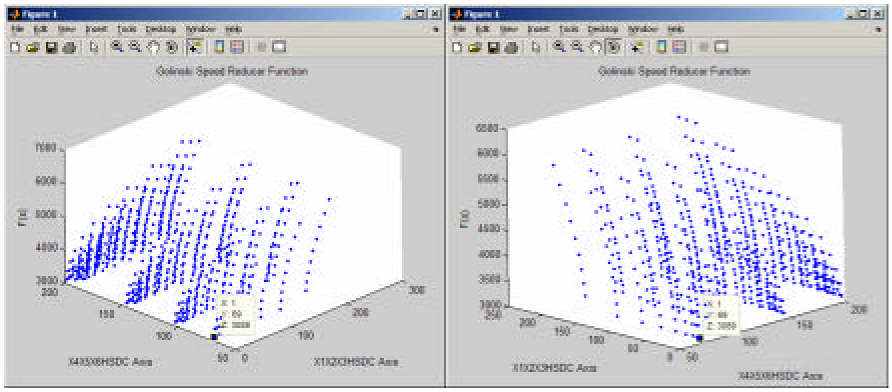


Figure 15. HSDC-based visual representation for speed reducer problem

This example problem demonstrates the potential use of the HSDC for either preprocessing prior to an expensive formal optimization or even replacement of an optimization altogether. Hypothetically, any number of variables can be blended on a single axis. Reference [31] has investigated blending of up to five variables per axis for MOP problems. Certainly, a more in-depth study of the viability of the HSDC for larger-scale problems is warranted.

V. HSDC's Application to Explore Multidimensional Datasets

Not only can the HSDC Method be used to explore multidimensional design spaces, but it can also be used equally well to highlight relationships or trends amongst existing high dimensional data. High dimensional datasets are used in everything from business to health care and engineering. The HSDC method allows for the bundling, or blending, of many dimensions into one contiguous, discrete space which is then mapped to an axis. Using a three dimensional representation, and assuming there are n dimensions to visualize, then some subset of dimensions can be mapped to the X axis, while the remaining dimensions can be mapped to the Y axis. It should be recognized that data comprising the type of dataset discussed here would not have an overarching functional relationship. Typically, data mining techniques would be used to answer questions of interest. Using the HSDC visualization approach, no explicit questions are asked at all. The data, such as it is, is presented in a hyperspace representation. The resulting representation can then be used to ask questions, rather than the reverse, as in data mining.

In order to be able to use the HSDC to visualize multidimensional datasets, we would need an index based approach to represent the data. Some metric, such as histogram, can then be plotted in the Z axis. The index based approach that has been developed to represent data using HSDC method is summarized below.

1) Identify minimum and maximum values for all the variables for the data to establish a range.
2) Divide these ranges into some finite number of compartments, resulting in small bins for each variable.
3) Group the variables into two sets and 'count' each set using HSDC, producing indices for each.
4) The indices of the bins can be plotted on a line and thus we can have indices for the bins of some variables on one axis and indices for the bins of some other variables on the other axis.
5) Determine what combination of indices corresponds to which bins that contains a data point. Each data point will fall under some combination of these bins that are plotted using their indices on the X and Y axes.
6) The points can be represented as a unit cylinder along the vertical axis. This will result in a 2-d histogram that represents all dimensions for the multidimensional data.
7) Multiple points might fall under the same set of indices, resulting in some bins that contain more than one data point.

The test cases discussed below demonstrate the use of HSDC along with the binning approach (to create an index based representation) to visualize multidimensional datasets.

A. Test Case 1 – Department of Labor Data

The following data set (Table 2) is readily available from the US Department of labor website (http://www.dol.gov). Four categories were downloaded for a five year period (1998-2003). The four attributes are CPI (Consumer Price Index), ML (Mass Layoffs), UN (Unemployment), and WS (Work Stoppage). Minimum and maximum values were identified within each of

Table 2. Department of Labor Data

CPI	ML	UN	WS	Year	Month
162	428	8.6	16	1998	1
162	154	8.7	38	1998	2
162	106	8.4	21	1998	3
162.2	189	8.2	0	1998	4
162.6	208	8	72	1998	5

these categories to establish a range for each. Each of these ranges (associated with the categories, which are the dimensions) was then subdivided into 50 compartments, each of which formed an index along that dimension. Fig. 14 shows the HSDC plot for the data. What is surprising about the visualizations in Fig. 16 is that we see two very clear locations (shown by data values in boxes) for which all four attributes are large simultaneously (where large is undesirable for these types of attributes).

When we investigate further, we discover that the two points in the figure on the left correspond to the time period November and December 2001. This occurs one and a half months following the attacks of September 11, 2001. The second set of points (in the figure on the right) correspond to the time period December 2002 and January 2003. This is one and a half months following the passing of the Iraq Resolution of War in the U.S. Congress. The country knew war was imminent and awaited only the final declaration by the President. The visualization is confirming something that, in retrospect, makes quite a bit of sense. That mass layoffs, unemployment, work stoppage, and consumer price index were all uniformly high during these periods of great strife in the United States – periods of time in which the U.S. economy was hard hit. The HSDC method, coupled with the visualization approach shown, was able to find these relationships without any bias. This is quite remarkable and demonstrates tremendous potential for investigating databases of nominal data. Recall that no explicit question was asked here, but rather the data was just visualized. By seeing the outliers associated with both axes simultaneously, the two sets of data discussed above were identified. One could also easily see which months would correspond to large values in CPI/UN and associated low values in ML/WS by merely following the x axis out to the maximum index. Other data can be explored equally well.

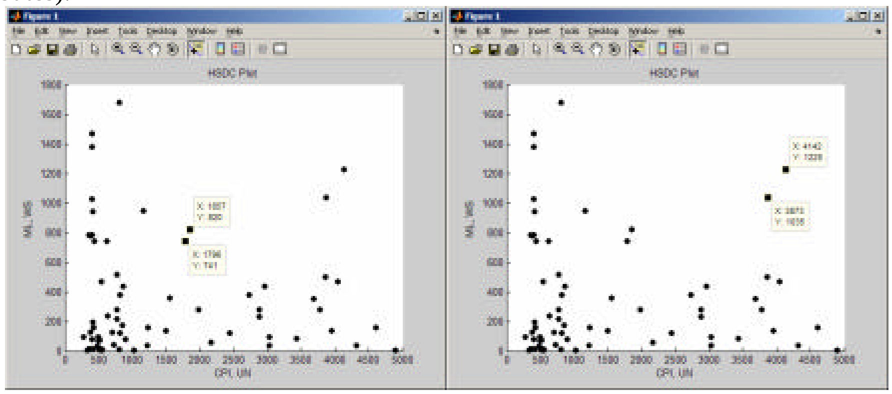


Figure 16. Two variate visualization of 4-d database

B. Test Case 2 – Public Use Micro Data

The second dataset (Table 3) used in our study was obtained from the United States Census Bureau website (http://www.census.gov). The data represents topcode values for various statistics about the housing and person variables. Specifically, they are ELEP (Electricity, monthly cost), GASP (Gas, monthly cost), WATP (Water, yearly cost), FULP (House Heating Fuel, yearly cost), RNTP (monthly rent), and MRGP (Mortgage Payment, monthly amount). The data is broken down into values for all the states. Each row represents one time-coincident set of values (year 2004). In other words, each row represents the values of the six indicators for a given state in the year 2004. In this example data set, we are interested in visually displaying what relationship these indicators have with each other. We have removed the state element from the data, and now simply want to plot the coincident ranges (i.e. how many times did these values fall into a specific set of ranges?).

Table 3. Public Use Micro Data Used for HSDC Data Mining

STATE	ELEP	GASP	WATP	FULP	RNTP	MRGP
AL	428	450	2431	4650	1458	3886
AK	457	460	3825	4676	1858	3855
AZ	464	372	3068	5718	2596	4865
AR	463	439	2178	2885	2142	3332
CA	516	367	3630	4152	3285	6922
CO	491	355	3054	2537	2123	4753

The first step again is to discretize each of the ranges of data. Once they've been discretized, we simply use the binning approach discussed previously to enumerate the possible combinations of different ranges for these variable and map them on axes using the HSDC method. We then walk the entire data set, shown in Table 3 (sample only), and keep a count for each of these combinations. Using the HSDC method, we blend the first three variables on the X axis (ELEP, GASP, and WATP) and remaining variables on the Y axis, and plot the histogram of all possible sub range combinations in Z. Fig. 17 shows the representation. The common attribute of all of these indicators is that higher numbers imply worse conditions. So, any points near the upper right corner indicate that all six of these values are toward the "high" end of the scale. The figure clearly shows a few outliers.

Upon further inspection of the data, the four outliers in Fig. 17 correspond to the states of Connecticut, New York, New Jersey, and Washington DC. The data shows that these states have relatively high values for these housing variables, reflecting a relatively high cost of living; where as the state of Idaho has the lowest overall values for these housing variables.

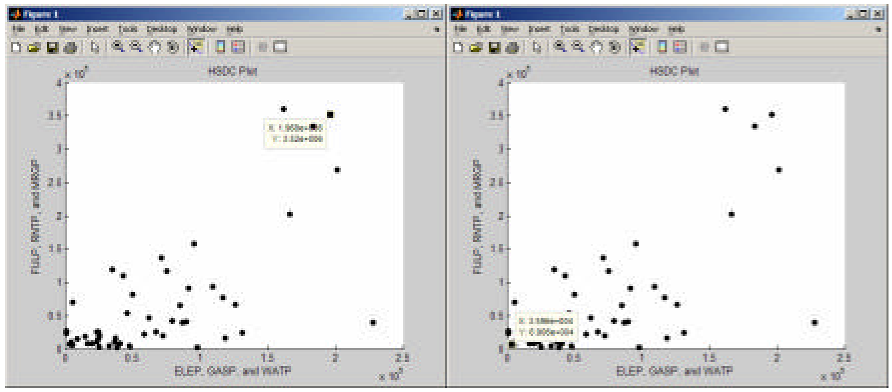


Figure 17. Two variate visualization of 6-d database

Here, if some of these variables were considered to be more important than others, a regrouping of the data could be easily implemented. The HSDC approach is very flexible in this regard. Again, as with the first test case, no questions were explicitly asked here.

C. Test Case 3 – Application to Multiobjective Optimization

Multiobjective Optimization Problems (MOPs), in contrast to single objective problems, involve a set of objectives that might be cooperative, competitive or have no relationship. Typically, the case of competitive objectives is the most interesting, since the choice of an ‘acceptable’ or ‘best’ solution depends on the preferences on, compromises between, and trade-offs of the objective functions. Most engineering design problems can be categorized as being multi-objective problems. When the objectives in an MOP are conflicting, the set of optimal solutions is known as the Pareto set, wherein no one solution is superior to the others. The region defined by the Pareto optimum is called the Pareto Front. Being able to visualize the Pareto frontier in a performance space with more than three objective functions has been a great challenge to the optimization community. The visualization of any performance space is currently limited to three dimensions (three objective functions). There are some other developments that integrate more than three dimensions using colors, shapes, glyphs, and other visual cues [29, 30, 31]. However, with these methods the users must be able to keep track of a sometimes complicated legend that correlates a visual cue to an attribute. Parallel coordinates [26, 27, 28] has been one of the leading approaches used in industry to assist in ultimately choose leading design candidates to explore further, but remains unwieldy for large numbers of functions.

With MOPs, the HSDC can be applied to visualize the performance space, just as we visualize any dataset (such as in the previous two test cases). Consider a four-objective problem[29, 30, 31] for which a total of 1384 Pareto points (i.e. design concepts) were obtained. A Hyperspace Pareto Frontier (HPF) [31], shown in Fig. 18, was generated by grouping the first and second

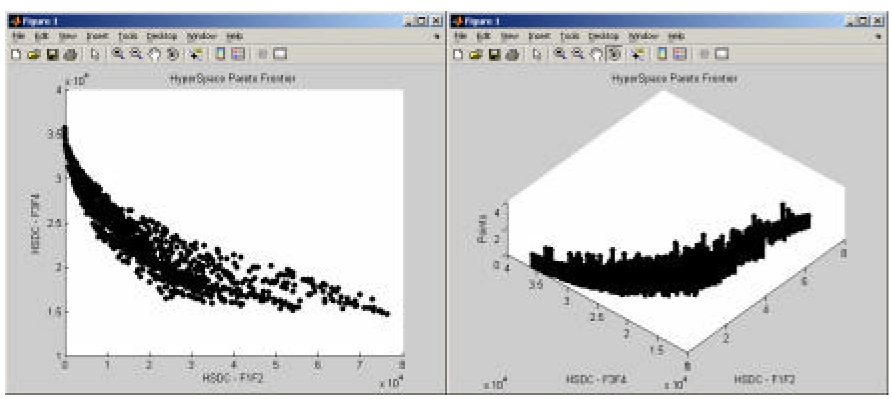


Figure 18. Hyperspace Pareto Frontier

objective on the X-axis and the third and fourth objective on the Y-axis, using the binning approach. The details of the HPF can be found in reference [31], which discusses the application of the HSDC to visualize the results of multiobjective optimization (resulting in the HPF representation). This reference also provides some strategies to incorporate a designer’s preferences into the representation, so as to allow an easy and efficient mechanism for concept selection in a multiobjective optimization environment. The HPF serves essentially the same purpose as a Pareto frontier for two objectives. Just as a two objective Pareto frontier can provide insights into design concepts as well as implicit trade-offs between functions, so can the HPF. The reader is encouraged to read reference [31] for a detailed description of the application of the HSDC to multiobjective optimization problems.

VI. Conclusion

In this paper, a visualization methodology is developed and presented that enables the visualization of multidimensional relationships in an intuitive and straightforward manner. The method presented is termed the *Hyper-Space Diagonal Counting (HSDC)* method for multidimensional visualization. It was demonstrated that the HSDC-based visualization method makes it possible to quickly identify 'good' regions of the multidimensional design space (for an optimization application) as well trends and outliers from multidimensional datasets, regardless of the complexity of the space or dataset. The HSDC method described in this paper enables a lossless approach to visualize multidimensional problems with more than three variables. The end result is an efficient and reliable mechanism to investigate relationships between multiple dimensions of problems with large number of variables, with minimal effort. The HSDC method described in this research requires no dimension fixing. The method overcomes the issues of awkwardness and dimension fixing that are so prevalent in the other existing visualization strategies. This is because the HSDC-based approach is a lossless representation, in which every single dimension is represented in a 2- or 3-D visual representation that is intuitive and easy to understand.

Acknowledgments

We gratefully acknowledge support from the New York State Center for Engineering Design and Industrial Innovation (NYSCEDII) and funding from NYSTAR (New York State Office of Science, Technology and Academic Research).

References

[1]Yang, J., Ward, M. O., Rundensteiner, E. A., and Huang, S., 2003, "Visual Hierarchical Dimension Reduction of High Dimensional Datasets," ACM International Conference Proceedings Series, Proceedings of the Symposium on Data Visualization

[2]Beddow, J., 1992, "An Overview of Multidimensional Visualization: Elements and Methods," Designing a Visualization Interface for Multidimensional Multivariate Data, IEEE Visualization '92 Tutorial 8, Boston, MA

[3]Wong, P. C., and Bergeron, R. D., 1997, "30 Years of Multidimensional and Multivariate Visualization," Department of Computer Science, University of New Hampshire, Durham, NH

[4]ATKOSoft S.A., 1997, "Survey on Visualization Methods and Software Tools," http://europa.eu.int/en/comm/eurostat/research/supcom.96/30/result/a/visualisation_methods.pdf

[5]Carreira-Perpinan, M. A., 1997, "A Review of Dimension Reduction Techniques," Technical Report CS 96-09, Department of Computer Science, University of Scheffield

[6]Friedman, J. H., and Tukey, J. W., 1974, "A Projection Pursuit Algorithm for Exploratory Data Analysis," IEEE Transactions Computers, C-23 (1974)

[7]Huber, P. J., 1985, "Projection Pursuit," The Annals of Statistics, Vol. 13, No. 2

[8]Kaiser, H. F., 1958, "The Varimax Criterion for Analytic Rotation in Factor Analysis," Psychometrika, Vol. 23

[9]Alsberg, P. A., 1975, "Space and Time Savings through Large Data Base Compression and Dynamic Restructuring," Proceedings of IEEE

[10]Fukunaga, K., 1990, "Introduction to Statistical Pattern Recognition," 2nd edition, Academic Press, New York

[11]Hardle, W., 1991, "Smoothing Techniques with Implementation in S," Springer Series in Statistics, Springer-Verlag, Berlin

[12]Scott, D. W., 1992, "Multivariate Density Estimation. Theory, Practice and Visualization," Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, London, Sydney

[13]Dunteman, G. H., 1989, "Principal Component Analysis," Sage University Paper Series on Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills

[14]Jackson, J. E., 1991, "A User's Guide to Principal Components," Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, London, Sydney

[15]Jolliffe, I. T., 2002, "Principal Component Analysis," 2nd edition, Springer Series in Statistics, Springer Verlag, Berlin

[16]Van Wijk, J. J., and Van Liere, R., 1993, "HyperSlice - Visualization of Scalar Functions of Many Variables," Proceedings of IEEE Visualization Conference, San Jose, CA

[17]Alpern, B., and Carter, L., 1991, "The HyperBox," Proceedings of IEEE Visualization Conference, San Diego, CA

[18]Chernoff, H., 1973, "The Use of Faces to Represent Points in k-Dimensional Space Graphically," Journal of American Statistical Association, Vol. 68

[19]Beddow, J., 1992, "An Overview of Multidimensional Visualization: Elements and Methods," Designing a Visualization Interface for Multidimensional Multivariate Data, IEEE Visualization '92 Tutorial 8, Boston, MA

[20]Beddow, J., 1990, 'Shape Coding of Multidimensional Data on a Microcomputer Display," Proceedings of IEEE Visualization Conference, San Francisco, CA

[21]Mihalisin, T., Gawlinski, E., Timlin, J., and Schwegler, J., 1990, "Visualizing a Scalar Field on an n-Dimensional Lattice," Proceedings of IEEE Visualization Conference, San Francisco, CA

[22]Mihalisin, T., Timlin, J., and Schwegler, J., 1991, "Visualization and Analysis of Multivariate Data: A Technique for All Fields," Proceedings of IEEE Visualization Conference, San Diego, CA

[23]LeBlanc, J., Ward, M. O., and Wittels, N., 1990, "Exploring n-Dimensional Databases," Proceedings of IEEE Visualization Conference, San Francisco, CA

[24]Feiner, S., and Beshers, C., 1990, "Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds," Proceedings of 3rd Annual ACM SIGGRAPH Symposium on User Interface Software and Technology, Snowbird, UT

[25]Feiner, S., and Beshers, C., 1990, "Visualizing n-Dimensional Virtual Worlds with n-Vision," Computer Graphics, ACM SIGGRAPH, Vol. 24, No. 2

[26]Inselberg, A., 1997, "Parallel Coordinates for Visualizing Multidimensional Geometry," Proceedings of New Techniques and Technologies for Statistics

[27]Inselberg, A., and Dimsdale, B., 1990, "Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry," Proceedings of IEEE Visualization Conference, San Francisco, CA

[28]Wegman, E. J., 1990, "Hyper-dimensional Data Analysis Using Parallel Coordinates," Journal of the American Statistical Association, Vol. 85, No. 411

[29]Agrawal, G., Lewis, K.E., Chugh, K., Huang, C-H., Parashar, S, and Bloebaum, C.L., 2004, "Intuitive Visualization of Pareto Frontier for Multi-Objective Optimization in n-Dimensional Performance Space," 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, AIAA-2004-4434

[30]Agrawal, G., Lewis, K.E., and Bloebaum, C.L., 2005, "Intuitive Design Selection Using Visualized n-Dimensional Pareto Frontier," 46th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference, Austin, TX, AIAA-2005-1813

[31]Agrawal, G., Lewis, K.E., and Bloebaum, C.L., 2006, "Intuitive Visualization of Hyperspace Pareto Frontier," 44th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV

[32]Golinski, J., 1970, "Optimal Synthesis Problems Solved by Means of Nonlinear Programming and Random Methods," Journal of Mechanisms