A Scatterplots Selection Technique for Multi-Dimensional Data Visualization Combining with Parallel Coordinate Plots

Ayaka Watanabe[1)], Takayuki Itoh[1)], Masahiro Kanazaki[2)], Kazuhisa Chiba[3)]
1) Ochanomizu University 2) Tokyo Metropolitan University,
3) The University of Electro-Communications
aya@itolab.is.ocha.ac.jp, itot@is.ocha.ac.jp, kana@tmu.ac.jp, kazchiba@uec.ac.jp

Abstract

We have previously presented a visualization technique which represents multi-dimensional data as collection of low-dimensional parallel coordinate plots. This paper presents a general-purpose extension of the visualization technique which represents multi-dimensional data as combination of the scatterplots with parallel coordinate plots. We aim to automatically select a small number of pairs of variables s which are estimated that they bring interesting visualization by scatterplots. This paper also presents an application of this visualization technique to Multi-objective optimization of manufacturing design. Our multi-dimensional data visualization technique effectively assists us to understand the distribution and correlation of design variables and objective functions in multi-objective optimization processes.

Keywords--- Multidimensional data, Visualization

1. Introduction

Multi-dimensional data visualization is an important issue in the field of information visualization. We previously presented a visualization method "Hidden" [1]. Using this method, we can represent correlation among dimensions by dimension scatterplots which places dimensions as dots. Also, we can visualize distributions of well-correlated sets of dimensions by Parallel Coordinate Plots (PCPs). We also presented an application of Hidden for an multi-objective optimization of airplane wing shape design [2].

Multi-objective optimization and data mining by visualizing the optimization results are often used in the field of airplane design. Selection of solutions get a complex task of multi-objective optimization, as the number of objective functions increases. Therefore, visualization of Pareto-optimum solutions is important.

We can treat both explanatory variables and objective functions of multi-objective optimization as multi-dimensional variables. Therefore, we can observe the distributions and correlations of datasets by using multi-dimensional visualization methods. As a study of visualization for airplane design, an interactive search of Pareto solutions by combination of explanatory variables and objective functions is proposed [3]. We can understand relationships between explanatory variables or between objective functions by using this method. However, it is difficult to observe relationships between explanatory variables and objective functions.

In this paper, we present an extension of Hidden by applying combination of scatterplots and low-dimensional PCPs. In this extension, we automatically select pairs of variables which are informative and worth visualizing by scatterplots. We select "informative" pairs of variables by their characteristic distributions, which are not effectively represented by PCPs but by scatterplots. On the other hand, PCP has an advantage of displaying many dimensions in a relatively compact screen space. This is the main reason we developed a method applying a combination of PCPs and scatterplots. This method applies scatterplots only when the distributions are not effective represented by PCPs.

This paper also presents an application of the method to visualization of multi-objective optimization results of manufacturing design. This paper introduces examples of scatterplots selected by our method.

2. Related work

Multi-objective optimization is an interesting application of visualization techniques which aim to assist understanding of the distribution of solutions. Eddy et al. [4] presented "Cloud Visualization" which displays Pareto-optimal solutions by scatterplots. Obayashi et al. [5] uses Self-Organizing Maps (SOMs) for clustering and visualization of Pareto solutions. However, these studies merely display distributions of objective functions; they did not implement visualizations of explanatory variables.

Scatter Plot Matrix (SPM) and PCP are most popular techniques for multi-dimensional data visualization. Even though many visualization techniques have been developed, it is not still always easy to observe complex relationships among variables of high-dimensional data by using these methods. We previously proposed visualization techniques [6,7] which automatically select and display PCPs and/or scatterplots which are worth

visualizing. However, these methods did not support a mechanism to interactively control PCPs and scatterplots.

Hidden [1] has been proposed to solve the above problems. Hidden features a user interface to semi-automatically select a set of low-dimensional subspaces consisting of well-correlated dimensions. Figure 1 is an example of visualization by Hidden. The left side of the window displays PCPs to show the low-dimensional subspaces. The right side displays scatterplots to visualize distances among dimensions, where dots in the scatterplots correspond to dimensions.

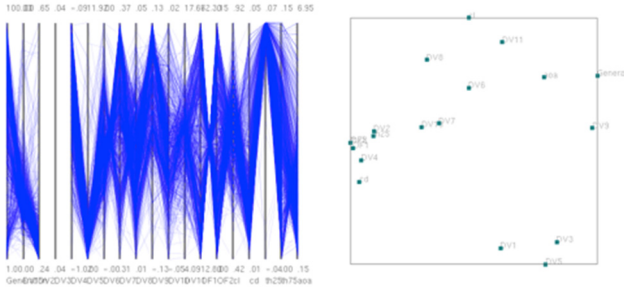

Figure 1 Example of multi-dimensional data visualization by Hidden [1]

3. Extended multi-dimensional data visualization by using both scatterplots and PCPs

This section describes a new extension of our multi-dimensional data visualization method Hidden. This extension automatically recommends pairs of variables which are worth visualizing by scatterplots. The original version of Hidden [1] displays PCPs in the left side of the drawing area. As discussed in Section 1, we visualize limited number of variables by using a combination of PCPs and scatterplots. We use scatterplots to pick up pairs of variables which have characteristic distributions, but are difficult to effectively represent by PCPs.

We apply "Graph-Theoretic Scagnostics" presented by Wilkinson et al. [8] to select the pairs of variables. This method quantifies nine types of features for each scatterplot. By using some of these measures, the presented extension recommend pairs of variables which are similar to one of the eleven types of representative scatterplots. We are currently applying two types of features: "Monotonic" and "Skinny".

3.1. Monotonic

"Monotonic" means monotonicity of a pair of variables. In other words, it means if the variable monotonically increases/decreases. Distribution of values can be effectively visualized by using PCPs is a pair of variables has a high Monotonic value. In this case, we do not need to apply scatterplots to visualize such pairs of variables. Therefore, our extension does not recommend pairs of variables which have high Monotonic values.

We adopt Spearman's rank correlation coefficient to calculate Monotonic value. Following is the formula of Spearman's rank correlation:

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \qquad \cdots\cdots(1)$$

Here, ρ is the coefficient, n is the total sample number and d_i is rank difference of the i-th pair of sample. We also apply average ranks if multiple samples have the same rank. This paper describes the threshold of Monotonic value as M_{select}.

3.2. Skinny

"Skinny" means thinness of a cloud of dots in scatterplots. If a cloud of dots is thin, one variable can be explained on the other variable with a certain degree of accuracy. Scatterplots must be worth visualizing, if a pair of variables has such a relationship. In our study, we select pairs of variables if they have high Skinny values, and display such pairs by scatterplots.

We apply Delaunay triangulation to evaluate "Skinny", as implemented by Wilkinson et al. [8]. Delaunay triangulation is a generic method which connects vertices scattered in a 2D space and generates a triangle mesh. We apply an incremental triangulation algorithm which firstly generates a big triangle surrounding every vertex in a scatterplot as the initial mesh, then adds the vertices one-by-one, and connects the vertices to refine the triangular mesh.

After constructing the mesh, our implementation deletes edges which are longer than a threshold S_{select}, and also deletes triangles which contain such edges. Here, let us describe the region consisting of the remaining triangle as A. We define the formula to calculate Skinny as follows:

$$Skinny = 1 - \sqrt{4\pi\, area(A)}/perimeter(A) \quad \cdots(2)$$

Here, $perimeter(A)$ denotes the length of outer boundary of A, and $area(A)$ denotes the area of A. Shape of A is thin if Skinny value is close to 1. We determine the scatterplot is worth visualizing if A is thin.

4. User interface

Figure 2 shows the user interface of our extended method. We place previous visualizations of Hidden on the right side of the drawing area, and scatterplots which visualize worth pairs of variables on the left side. In addition, we specify the priority of scatterplots by the following steps:

1. Compute Monotonic values for each pair of variables (corresponds to each scatterplot).
2. If an absolute value of Monotonic is larger than the threshold (M_select), this plot is classified as "high priority" because the relationship between the pair of variables is difficult to be observed by a PCP. If smaller, this plot is classified as "low priority".
3. Sort scatterplots in the order of Skinny values.

We can interactively switch the scatterplots to be displayed by using sliders or buttons featured on the left side of the window. Users can select any pairs of variables, by entering numbers of variables and pressing "Dataset" button. Then, the corresponding scatterplots are displayed on the left side of the drawing area.

We can select scatterplots based on the order of their priority by operating the slider. The right end of the slider corresponds to the 1st place scatterplots, and the left end to the last place. The scatterplots corresponding to the position of slider is displayed on the drawing area, when a user operates the slider. We can also select scatterplots by directly entering their orders and pressing "NumOfSP (Number Of Scatterplot)" button. Moreover, users can switch scatterplots one-by-one, when a user presses plus or minus button in line with "NumOfSP" button, where plus denotes the next higher place, and minus denotes the next lower place.

Moreover, users can adjust the thresholds of Monotonic and Skinny by using two boxes at the bottom of the window. Other sliders and boxes are reset when they use the two boxes.

We developed above functions consistently so that any set of sliders and boxes do not contradict after operations.

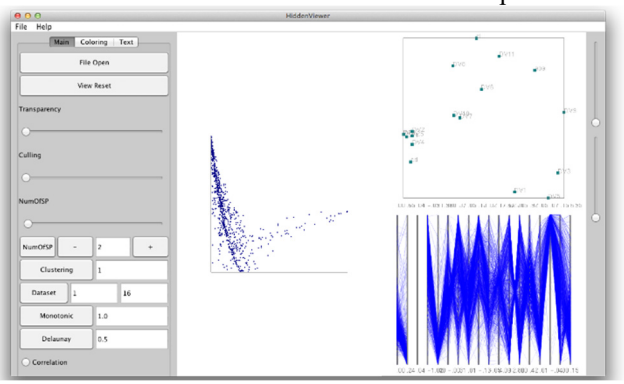

Figure 2 Snapshot of the extended Hidden

5. Application example

Our method are implemented by Java Development Kit (JDK) 1.7.0. In this chapter, we summarize the data of airfoil design for a Martian airplane [9] and hybrid rocket design [10].

5.1. Airfoil of Martian airplane

We applied Hidden to design optimization results for an airfoil of Martian airplane. The problem is to design an airfoil that achieves a high glide ratio and higher structural performance under the unknown conditions of the Martian atmosphere.

Objective functions are as follows:

- OF1:to maximize the maximum lift-to-drag ratio (*maxl/d*)
- OF2:to maximize thickness at 75% chord length (*th75*.)

Design variables are as follows:

- Thickness distribution.
  - DV1:radius of the leading edge
  - DV2:x-coordinate of maximum thickness
  - DV3: maximum thickness
  - DV4: curvature at maximum thickness
  - DV5: angle of aperture of trailing edge
- Camber distribution.
  - DV6: curvature of chamber at leading edge
  - DV7:x-coordinate of maximum chamber
  - DV8: maximum chamber
  - DV9: curvature at maximum chamber
  - DV10: chamber angle at trailing edge
  - DV11: y-coordinate of trailing edge

We applied a genetic algorithm to these variables and conducted evolutionary calculation with 10 individuals and 100 generations. As a result, we calculated 1000 solutions, and visualized them as a set of 13-dimensional vectors (11 explanatory variables + 2 objective functions).

5.2. Hybrid rocket

We applied Hidden to design optimization results for a hybrid rocket.

Objective functions are as follows:

- the maximization of the down range in the lower thermosphere
- the maximization of the duration time in the lower thermosphere Td [sec]
- the minimization of the initial gross weight of launch vehicle Mtot (0) [kg]

Design variables are as follows:

- DV1: initial mass flow of oxidizer
- DV2: fuel length
- DV3: initial radius of port
- DV4: combustion time
- DV5: initial pressure in combustion chamber
- DV6: aperture ratio of nozzle
- DV7: elevation at launch time

We visualized solutions as a set of 10-dimensional vectors (7 explanatory variables + 3 objective functions).

6. Visualization of low-dimensional space by scatterplots

We calculated Monotonic and Skinny of all possible pairs of variables of the two datasets introduced Section 5. The results are sorted and listed in Figures 3 to 6. In these figures, Monotonic or Skinny values are assigned to the vertical axis, and ranks of the variable pairs are assigned to the horizontal axis.

6.1. Airfoil of Martian airplane

As shown in Figure 3, Monotonic values are widely distributed. On the other hand, Figure 4 illustrates that Skinny values are concentrated on the range from 0.7 to 0.8. These results denote that most of scatterplots generated from pairs of variables have a "thin" field of points and are worth visualizing. In other words, it is difficult to select small number of important scatterplots based on Skinny values.

Range of values are below.

- Monotonic :
  -0.98486～0.99999
- Skinny :
  0.52801～0.98291

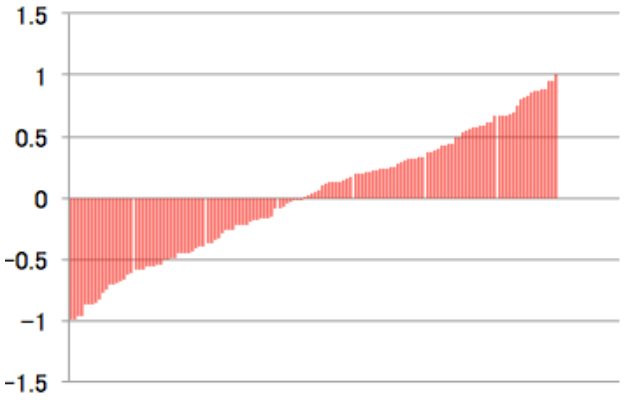

Figure 3 Monotonic values of every possible pairs of variables in the dataset of airfoil of Martian airplane

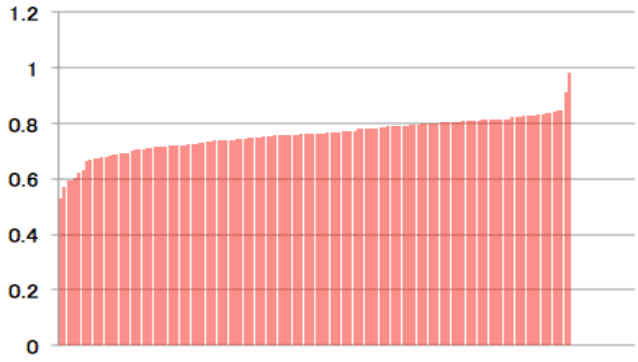

Figure 4 Skinny values of every possible pairs of variables in the dataset of airfoil of Martian airplane

6.2. Hybrid rocket

As shown in Figures 5 and 6, both Monotonic and Skinny values are widely distributed. Especially, Figure 6 shows that scatterplots have a variety in their thinness of a field of points comparing with Figure 4.

Range of values are below.

- Monotonic:
  -0.90953～0.98071
- Skinny:
  0.35603～0.92343

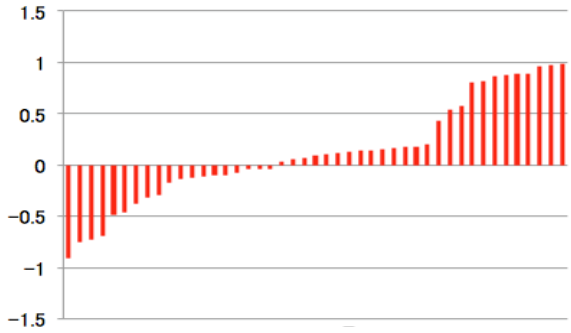

Figure 5 Monotonic values of every possible pairs of variables in the dataset of airfoil of hybrid rocket

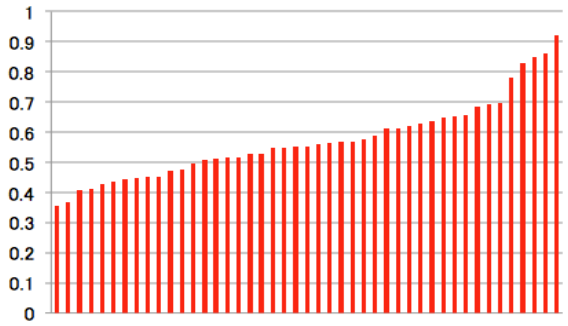

Figure 6 Skinny values of every possible pairs of variables in the dataset of airfoil of hybrid rocket

7. Examples of recommended scatterplots

This section introduces examples of scatterplots recommended by our technique. Here, we set threshold of Monotonic (M_{select}) as 0.8, and that of Skinny (S_{select}) as 0.6.

7.1. Airfoil of Martian airplane

From the design optimization result of airfoil of Martian airplane introduced in Section 5.1, the pair lift coefficient and drag coefficient was selected as the 1st place, as shown in Figure 7. The 2nd place was the pair DV2 and OF2, as shown in Figure 8. Though these two scatterplots have different appearances, both scatterplots are “Skinny”, and difficult to understand the distributions only by using PCPs.

7.2. Hybrid rocket

From the design optimization result of hybrid rocket introduced in Section 5.2, the pair DV1 and DV6 was selected as the 1st place, as shown in Figure 9. The 2nd to 4th place scatterplots also had similar appearances as Figure 9. The 5th place scatterplot was the pair DV2 and DV3, as shown in Figure 10.

From the above results, recommended scatterplots are “Skinny”, but do not have higher Monotonic values.

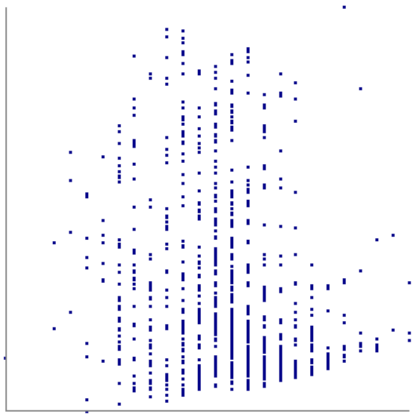

Figure 7 The 1st place scatterplots in the dataset of airfoil of Martian airplane

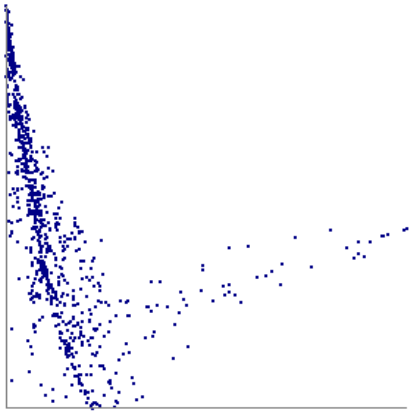

Figure 8 The 2nd place scatterplots in the dataset of airfoil of Martian airplane

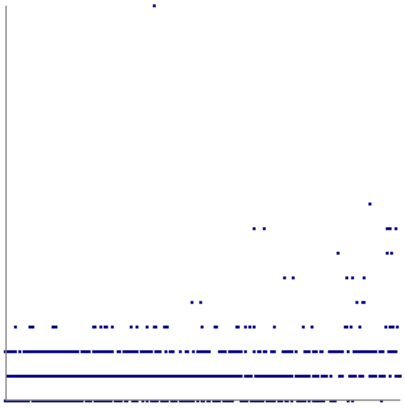

Figure 9 The 1st place scatterplots in the dataset of airfoil of hybrid rocket

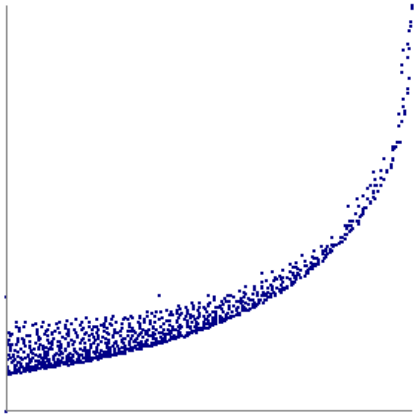

Figure 10 The 5th place scatterplots in the dataset of airfoil of hybrid rocket

8. Conclusions

This paper presented an extension of our multi-dimensional data visualization method "Hidden" which mixes PCPs and scatterplots to visualize well-correlated sets of dimensions. Also, the paper presented an application of this visualization technique for optimizing airplane design. Specifically, we developed a method to determine which pairs of variables are worth visualizing by scatterplots. We applied two criteria for evaluating scatterplots: "Monotonic" and "Skinny". We also developed a user interface to interactively adjust thresholds for the criteria.

Furthermore, we applied the multi-objective optimization results for design of Airfoil of Martian airplane and Hybrid rocket. In these examples, we set M_{select} (threshold of Monotonic) as 0.8 and S_{select} (threshold of Skinny) as 0.6. We presented the results to demonstrate the effectiveness of our method.

As future work, we will implement other criteria in addition to Monotonic and Skinny. Also, we would like to more tightly connect Hidden with evolutionary computation such as genetic algorithm.

References

[1] T. Itoh, A. Kumar, K. Klein, J. Kim, High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots, arXiV preprint, 1609.05268, 2016.

[2] A. Watanabe, T. Itoh, M. Kanazaki, M. Utsugi, K. Chiba, Multidimensional data visualization for airplane design optimization, DEIM Forum 2016 F4-3, 2016.

[3] M. Kubota, T. Itoh, S. Obayashi, Y. Takeshima, EVOLVE: A Visualization Tool for Multi Objective Optimization Featuring Linked View of Explanatory Variables and Objective Functions,18th International Conference on Information Visualisation (IV2014), pp. 351-356, 2014.

[4] J. Eddy, K. Lewis, Visualization of Multidimensional Design and Optimization Using Cloud Visualization, ASME Design Engineering Technical Conferences, DETC02/DAC-2006.

[5] S. Obayashi, D. Sasaki, Visualization and Data Mining of Pareto Solutions Using Self-Organizing Map, Lecture Notes in Computer Science 2632: Evolutionary Multi-Criterion Optimization 2003, pp. 796-809.

[6] H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Arrangement of Low Dimensional Parallel Coordinate Plots for High Dimensional Data Visualization, 17th International Conference on Information Visualisation (IV2013), pp. 59-65, 2013.

[7] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, Y. Kawahara, Scatterplot Layout for High Dimensional Data Visualization, Journal of Visualization, Vol. 18, No. 1, pp. 111-119, 2015.

[8] A. Wilkinson, R. Anushka, L. Grossman, Graph Theoretic Scagnostics, IEEE Symposium on Information Visualization, pp. 21-28, 2005.

[9] M. Utsugi, M. Kanazaki, T. Sato, K. Matsushima, Multi-ObjectiveDesign of Airfoil for Martin Airplane considering Trailing Edge Thickness, 30th International Symposium on Space Technology and Science, Kobe, Japan, July, 2015.

[10] K. Chiba, M. Kanazaki, M. Nakamiya, K. Kitagawa, and T. Shimada, Diversity of Design Knowledge for Launch Vehicle in View of Fuels on Hybrid Rocket Engine,Journal of Advanced Mechanical Design, Systems, and Manufacturing, Vol.8, No.3, p.JAMDSM0023, pp.1-14, 2014