Scatterplot Selection Applying a Graph Coloring Algorithm

Takayuki Itoh
itot@is.ocha.ac.jp
Ochanomizu University
Bunkyo, Tokyo, Japan

Asuka Nakabayashi
asuka@itolab.is.ocha.ac.jp
Ochanomizu University
Bunkyo, Tokyo, Japan

Mariko Hagita
hagita@is.ocha.ac.jp
Ochanomizu University
Bunkyo, Tokyo, Japan

ABSTRACT

Scatterplot selection is an effective approach to represent essential portions of multidimensional data in a limited display space. Various metrics for evaluating scatterplots, such as scagnostics, have been devised and applied to scatterplot selection. This paper presents a new scatterplot selection technique that applies multiple metrics. First, the technique calculates the scores of scatterplots with multiple metrics and then constructs a graph by connecting similar scatterplots. Next, it uses a graph coloring algorithm to assign different colors to similar scatterplots. We can extract a set of various scatterplots by selecting them that the specific same color is assigned. This paper introduces a visualization example with a retail dataset containing multidimensional climate and sales values.

KEYWORDS

Scatterplot selection, Graph coloring algorithm, Multidimensional data visualization.

ACM Reference Format:
Takayuki Itoh, Asuka Nakabayashi, and Mariko Hagita. 2021. Scatterplot Selection Applying a Graph Coloring Algorithm. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

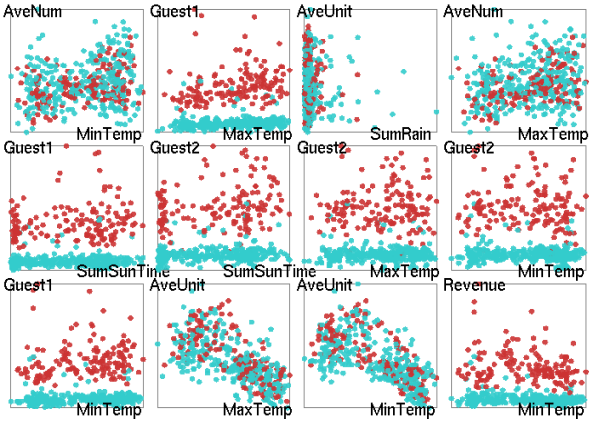

Figure 1: Visualization example by our technique. Several scatterplots show strong correlations between dimension pairs; some scatterplots clearly show clusters or outliers; several scatterplots show how two labels drawn in red and blue are separated. Our technique selects various scatterplots to show the various characteristics of the input multidimensional dataset in a single display space.

Multidimensional data visualization has been an active research area in the visualization community. Dimension selection [5, 18, 20] is a hotspot for visualizing high-dimensional data. It is unreasonable to represent every dimension in a limited display space; therefore, it is essential to remove noisy or meaningless dimensions and focus on visualizing informative dimensions.

Scagnostics [17] is a set of typical metrics applied to scatterplot selection problems [9, 16, 21]. We can selectively display a set of similarly featured scatterplots by applying one of the metrics. Meanwhile, it is not always suitable to apply a single metric for a scatterplot selection to capture all the characteristics of multidimensional data. For instance, interesting correlations are observed from some pairs of dimensions, while interesting clusters are observed from some other pairs of dimensions. We need operations to switch the metrics to display various types of scatterplots in this case. In other words, it is reasonable to select scatterplots that have various characteristics by applying multiple metrics simultaneously and displaying them in a single display space to understand various characteristics of the dataset without operations to switch the metrics.

This paper presents a new and fast technique for scatterplot selection with multiple metrics. First, this technique generates scatterplots with arbitrary pairs of dimensions. Then, it calculates multiple scores based on multiple metrics for each scatterplot and forms a vector from the scores. Next, it constructs a graph by connecting pairs of scatterplots if it determines that at least one of them can be eliminated. Further, it assigns colors to the vertices corresponding to the scatterplots while complying with a rule that different colors are assigned to a pair of vertices connected by an edge. In other words, the same color is assigned to a set of significantly different scatterplots. The technique selectively displays a constant number of scatterplots that have the same color. As shown in Figure 1, the technique realizes the selection of various scatterplots that show several characteristics of the input dataset.

This paper introduces a case study with a consumer business dataset, including climate and revenue values.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 RELATED WORK

1.1 Dimension Selection for Multidimensional Data Visualization

Dimension selection techniques have been widely applied to multidimensional data visualization to effectively represent essential subsets of dimensions. Claessen et al. [2] visualized high-dimensional datasets by representing a set of low-dimensional subspaces as a combination of PCPs (parallel coordinate plots) and scatterplots. Suematsu et al. [14] and Zheng et al. [21] also converted high-dimensional datasets into low-dimensional subsets and visualized these subsets using multiple PCPs or scatterplots, respectively. These techniques did not provide rich interaction mechanisms to freely select the numbers of dimensions.

Several recent studies have demonstrated interaction mechanisms to freely visualize interesting low-dimensional subspaces. Lee et al. [6] and Liu et al. [7] applied dimension reduction schemes to interactively select subsets of high-dimensional data. Nohno et al. [10] presented a technique to interactively contract highly correlated dimensions to adjust the number of axes displayed in PCPs. Itoh et al. [5], Watanabe et al. [16], and Nakabayashi et al. [9] presented a series of techniques that easily control the number of dimensions displayed in the PCPs or the number of dimension pairs represented by scatterplots.

It is also important to understand relationships among dimensions while extracting low-dimensional subspaces. Dimension spaces have been visualized by applying scatterplots or graphs by several recent studies [5, 18, 19]. This is an effective approach to interactively select reasonable sets of dimensions.

Despite many studies on multidimensional data visualization employing dimension selection techniques, there have been few studies to automatically select various limited number of informative scatterplots. We address this problem and present a new technique in this paper.

1.2 Evaluation of Scatterplots

Numeric evaluation of the informativeness of scatterplots has been an active research topic. Scagnostics is a remarkable method to quantitatively evaluate the informativeness of scatterplots. Wilkinson et al. [17] proposed nine features of scagnostics based on the appearance of scatterplots. Wang et al. [15] proposed an improved scagnostics by considering the human perception to several metrics, including "Outlying" and "Clumpy." There have been several more studies that focus on specific metrics of scatterplots, including correlation [4, 12] and class separation [1, 11, 13].

There have been several visualization studies on the overview and exploration of a large number of scatterplots. Dang et al. [3] presented an exploration mechanism for finding similar scatterplots and filtering scagnostics. Matute et al. [8] presented another approach to represent the distribution of characteristics of scatterplots. The goal of our study is somewhat similar to the above studies since we also focus on representing various scatterplots; however, our focus is different from these studies in that we aim to selectively display the user-defined number of various scatterplots.

2 SCATTERPLOT SELECTION APPLYING A GRAPH COLORING ALGORITHM

This section presents a processing flow of the presented scatterplot selection technique. The technique calculates the scores of scatterplots based on multiple metrics and stores it as vector values. Figure 2 illustrates the concept of scatterplot selection. Scatterplots are depicted as vectors in the metric space. The requirements for scatterplot selection in this study are summarized as follows.

R1: Avoid selecting sets of close vectors to avoid selecting similar scatterplots.

R2: Avoid selecting short vectors to avoid selecting less informative scatterplots.

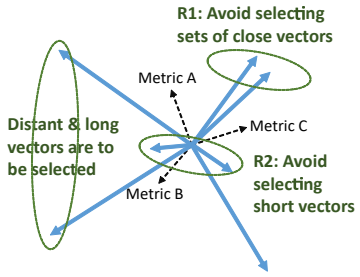

Figure 2: Concept of scatterplot selection in the metric space. Blue arrows illustrate the vectors of metrics. Our technique selects various scatterplots while satisfying R1 and R2.

The technique applies a graph coloring algorithm to satisfy the above requirements and displays various informative scatterplots.

2.1 Data Structure

This paper formalizes the problem as follows. An input multi 測 dimensional dataset A has n individuals as $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n\}$. The i-th individual $\boldsymbol{a}_i$ has the m-dimensional values as $\boldsymbol{a}_i = (a_{i1}, a_{i2}, ..., a_{im})$. A set of scatterplots formed from arbitrary pairs of dimensions is described as $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_N\}$, where N is the total number of scatterplots. Each scatterplot has a set of scores calculated based on predefined metrics, described as $\boldsymbol{s}_i = (s_{i1}, s_{i2}, ..., s_{iM})$, where M is the number of metrics.

2.2 Graph Coloring Algorithm

This technique applies a graph coloring algorithm to select various scatterplots with different characteristics. Suppose a graph $G = \{S, E\}$, where S is a set of vertices corresponding to the scatterplots, and E is a set of edges connecting pairs of scatterplots. Here, we select a pair of distant and long vectors, as shown in Figure 2. In other words, we would like to select a pair of the i-th and j-th vectors if the area of the triangle d_{ij} constructed by these vectors is large. Here, the technique constructs the graph by generating edges between the i-th and j-th scatterplots if the area of the triangle d_{ij} is smaller than the predefined threshold d_{thres}.

Then, the technique assigns colors to the scatterplots while complying to a rule that different colors are assigned to a pair of vertices connected by an edge. In other words, the same color is assigned to a set of significantly different scatterplots. Figure 3 illustrates the process. First, the process selects the scatterplot that has the largest $|s_k|$ and assigns the color identification $c_k = 0$. Then, adjacent vertices connected by edges are traversed in the breadth-first order. While visiting the k-th vertex, the process specifies the minimum color identification that is assigned to none of the adjacent vertices connected with the k-th vertex and assigns it to the k-th vertex. For instance, if color identifications 0, 1, and 3 have been assigned to the vertices adjacent to c_k, the process specifies c_k as 2. The breadth-first search is repeated until color identifications are assigned to all vertices.

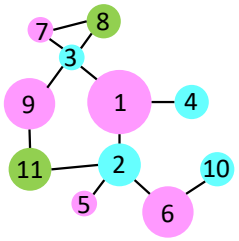

Figure 3: Graph coloring. The process assigns different colors to the vertex pairs connected by edges. The numbers in this figure denote the order of the breadth-first search.

Finally, we select a predefined number of scatterplots to be displayed. The technique extracts a set of scatterplots in which the same color is assigned. We calculate the sums of the length of the vectors $|s_k|$ for each color and select the color that brings the largest sum. The extracted set of scatterplots excludes similar or less informative pairs because such pairs of scatterplots are connected and therefore have different colors. In other words, it satisfies **R1** because the extracted set comprises various differently looking scatterplots. If the number of extracted scatterplots is larger than the user-defined number, the technique selects the scatterplots in descending order of $max(s_{ij})$, the maximum value of the scores of each scatterplot, to satisfy **R2**. Our implementation provides a user interface to interactively specify the number of scatterplots to be displayed.

The processing flow is as follows.

(1) Initialize the vertices S and calculate the interestingness of the k-th scatterplot as $|s_k|$.
(2) Construct the graph and generate an edge between the i-th and the j-th scatterplots if d_{ij} is larger than the pre-defined threshold.
(3) Select the scatterplot that has the largest interestingness as the starting vertex.
(4) Traverse the connected vertices by the breadth-first search. Assign color identifications to the traversed vertices. Repeat this traverse until the color identifications are assigned to all the vertices.
(5) Collect the vertices that have the same color identification. Select the user-defined number of vertices in the descending order of the maximum value of the scores of each scatterplot.

The problem solved using the above algorithm is similar to the maximum independent set problem. The presented algorithm is better for our study because it prioritizes to select "long" vectors and "distant" vectors.

3 APPLICATION TO RETAIL TRANSACTION DATA

This paper introduces an example of visualization by the presented technique applying a retail transaction and climate dataset. Table 1 shows the explanatory variables (climate values) assigned to the horizontal axis and the objective functions (retail transaction values) assigned to the vertical axis in this dataset. We clarified how the retail transaction values can be estimated from the climate values by visualizing them. The dataset contained the records of 457 days from May 1, 2016, to July 31, 2017, corresponding to 457 data points in the scatterplots. We generated 35 scatterplots consisting of five horizontal axes and seven vertical axes. The data points are drawn in red or blue; red denotes holidays, while blue denotes weekdays.

Table 1: The explanatory variables and the objective functions.

explanatory variables (climate values)	
MinTemp	Minimum temperature
MaxTemp	Maximum temperature
SumRain	Precipitation
SumSunTime	Sunshine duration
MaxWind	Maximum wind speed

objective functions (retail transaction values)	
Revenue	Revenue
Guest1	Number of customer
Guest2	Number of visitor
Ratio	Conversion rate
PerGuest	Average revenue per customer
AveUnit	Average price of purchased items
AveNum	Average number of purchased items

3.1 Selection of Metrics

Based on the discussion with data owners, we focused on finding the following scatterplots.

S1: Scatterplots with the variables that can contribute to the regression for predicting transaction values from climate values.
S2: Scatterplots representing outliers or isolated clusters.
S3: Scatterplots that separate different attributes (e.g., weekdays and weekend) of the plots.

We implemented the following four metrics to assist finding the above scatterplots.

3.1.1 Correlation. Correlation is one of the most common metrics used to determine the relationship between a pair of dimensions. It is an effective metric used to find tightly-correlated pairs of variables and find **S1**. Our current implementation just calculates the score of the k-th scatterplot as follows:

$$s_{k1} = |S_{pear}(i,j)^2| \tag{1}$$

where $S_{pear}(i,j)$ is the Spearman's rank correlation between the i-th and j-th dimensions. A dimension pair gets a higher score if they have a strong positive/negative correlation. Newer approaches on correlation [4, 12] can also be applied.

3.1.2 Thinness. It is easier to adopt a mathematical model to a set of individuals if they form thin regions in a scatterplot. Such scatterplots correspond to **S1**. We measure the thinness of the region where individuals are placed in the scatterplot as Wilkinson et al. [17] did. Our implementation generates a Delaunay triangular mesh T connecting the individuals in a scatterplot and then removes all triangles that have at least one edge that is longer than a predefined threshold. Then, we calculate the score as follows:

$$s_{k2} = 1 - \sqrt{4\pi A_{rea}(T)}/P_{erimeter}(T) \tag{2}$$

where $A_{rea}(T)$ is the total area of T, and $P_{erimeter}(T)$ is the total length of the boundary of T.

3.1.3 Clumping. It is remarkable if the individuals in a scatterplot are well-separated into several outliers and clusters. Such scatterplots correspond to **S2**. Our current implementation simply applies the metric "Clumpy" presented by Wilkinson et al. [17] defined as follows:

$$s_{k3} = 1 - length(e_{maxr})/length(e_{mind}) \tag{3}$$

Here, our implementation generates a Delaunay triangular mesh, as described in the previous section, and deletes the edges longer than e_{mind}. e_{maxr} is the longest remaining edge. Newer approaches on clumping [15] can also be applied.

3.1.4 Separateness. Suppose that one of the labels is assigned to each individual. It is remarkable if individuals that have a particular same label are well-separated in a scatterplot. Such scatterplots correspond to **S3**. We measure the separateness of a particular label by calculating the entropy of the labels. Particularly, we compute the entropy of the labels in the scatterplot generated with the i-th and j-th dimensions as follows:

$$H(i,j) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{C} p(y_k = c|(a_{ki}, a_{kj}) \log p(y_k = c|(a_{ki}, a_{kj})) \tag{4}$$

where y_k is the label of the k-th individual, (a_{ki}, a_{kj}) is the position in the scatterplot of the k-th individual, and C is the number of labels. Our implementation divides the scatterplot into L subareas and calculates the entropy at the l-th subarea $H(i,j)_l$ using the above equation, and finally calculates the score of the k-th scatterplot as follows:

$$s_{k4} = (H_{max} - \sum H(i,j)_l)/H_{max} \tag{5}$$

where H_{max} is the maximum value of $\sum H(i,j)_l$.

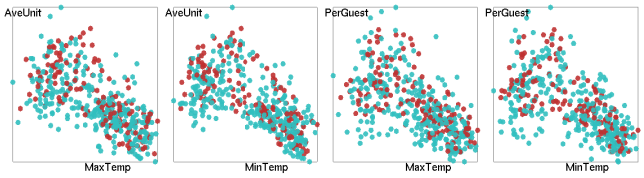

Figure 4: Scatterplots that achieved the highest scores on correlation.

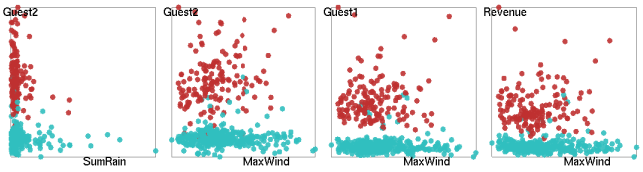

Figure 5: Scatterplots that achieved the highest scores on separateness.

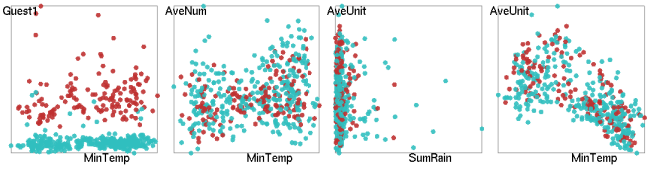

Figure 6: Scatterplots that achieved the highest scores on clumpy.

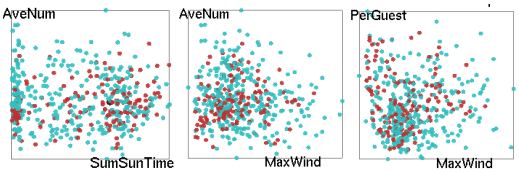

Figure 7: Example of scatterplots that have no higher scores with all metrics.

Other approaches [1, 11] can be also applied to determine the class separateness.

3.2 Results

Figure 1 shows an example of a scatterplot selection using our technique. Here, several scatterplots show correlations between dimension pairs, some others show clusters or outliers, while several others show how two labels drawn in red and blue are separated. This figure demonstrates that our technique successfully selects various scatterplots to show various characteristics of the dataset.

Figures 4, 5, and 6 show top four scatterplots that achieved the highest scores on correlation, separateness, and clumpy. The horizontal axes of scatterplots are MinTemp or MaxTemp, while the vertical axes are PerGuest or AveUnit (Figure 4). This implies that the average revenue or price correlates well with the temperature.

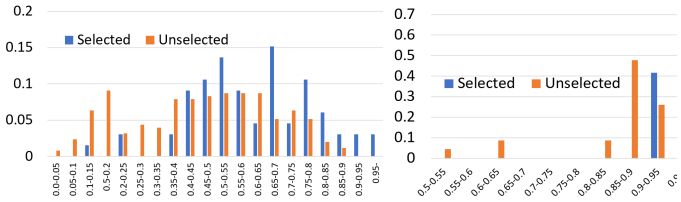

Figure 8: Statistics of scatterplots selected/unselected in Figure 1. Vertical axes denote the ratios of the number of corresponding scatterplots. (Left) Statistics of areas d_{ij}. (Right) Statistics of maximum score values $max(s_{ij})$.

Meanwhile, the vertex axes of scatterplots in Figure 5 are Revenue, Guest1, or Guest2. It implies that revenue and the number of guests significantly differ between holidays and weekdays. Figure 6 shows a set of dimension pairs that bring better views to discover outliers and clusters. The scatterplot selection result shown in Figure 1 is well-balanced because it represents various characteristics of the input dataset by selecting various scores of scatterplots. Meanwhile, Figure 7 shows examples of scatterplots that have no higher scores with all metrics. These scatterplots do not look characteristic or informative. The presented technique does not aggressively select such scatterplots.

Figure 8 shows the statistics of areas d_{ij} and maximum score values $max(s_{ij})$ of the scatterplot selected/unselected in Figure 1. This figure demonstrates that the presented technique tends to select scatteplots that have larger $max(s_{ij})$ values and pairs of scatteplots that have larger d_{ij} values preferentially.

The result of scatterplot selection strongly depends on the choice of d_{thres}. The smaller d_{thres} brings a larger number of edges and consequently a larger number of scatterplots groups corresponding to the number of colors in Figure 3. Table 2 shows the numbers of edges and colors, the number of scatterplots belonging to the selected color. The table also shows the minimum d_{ij} and $max(s_{ij})$ values among the displayed scatterplots supposing twelve of them are displayed. The result shows that selection of very similar or less informative scatterplots would be avoided when a larger number of colors are made and the minimum d_{ij} value gets larger. But simultaneously, the informativeness of the selected scatterplots may be decreased because the minimum $max(s_{ij})$ values get smaller. In other words, one of the features of the presented technique is that users can easily deal with this trade-off problem just by adjusting the d_{thres} value.

Table 2: Trade-off between the minimum d_{ij} and $min(s_{ij})$ values.

d_{thres}	0.45	0.5	0.55	0.6	0.65
Num. edges	26	49	124	228	299
Num. colors	7	9	10	14	19
Num. scatterplots	29	26	26	20	15
minimum d_{ij}	0.5158	0.5158	0.5549	0.6115	0.6512
minimum $max(s_{ij})$	0.9289	0.9289	0.9247	0.9189	0.8935

4 CONCLUSION AND FUTURE WORK

This paper presented a new scatterplot selection technique using a graph coloring algorithm. The technique calculates scores based on several independent metrics for each scatterplot. Then, it constructs a graph by connecting vertex pairs corresponding to scatterplot pairs if these scores are similar. The graph coloring algorithm is used for the graph, and scatterplots that have the user-specified color are extracted. The paper introduced an example of the scatterplot selection applying a retail transaction and climate dataset.

Our future studies include the following. First, we would add and modify the metrics. There have been various improved metrics for scagnostics. We will apply them and explore the best combination of the metrics for this study. Then, we will test the scalability of the presented technique. Particularly, we suppose it is necessary to test datasets with a large number of dimensions; therefore, a large number of scatterplots can be generated. In addition, it is necessary to test datasets with a large number of individuals. After the above improvements and tests, we will consider case studies with various real-time datasets and conduct user evaluations.

REFERENCES

[1] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *IEEE Pacific Visualization Symposium 2012*, pages 43–52, 2016.
[2] J. H. T. Claessen and J. J. van Wijk. lexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2310–2316, 2011.
[3] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *IEEE Pacific Visualization Symposium 2014*, pages 73–80, 2014.
[4] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber ' s law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.
[5] T. Itoh, A. Kumar, A. Klein, and J. Kim. High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. *Journal of Visual Languages and Computing*, 43(1):1–13, 2017.
[6] J. H. Lee, K. T. McDonell, A. Zelenyuk, D. Imre, and K. Muller. A structure-based distance metric for high-dimensional space exploration with multidimensional scaling. *IEEE Transaction on Computer Graphics*, 20(3):351–364, 2013.
[7] S. Liu, B. Wang, P.-T. Bremer, and V. Pascucci. Distortion-guided structure-driven interactive exploration of high-dimensional data. *Computer Graphics Forum*, 33(3):101–110, 2014.
[8] J. Matute, A. C. Telea, and L. Linsen. Skeleton-based scagnostics. *IEEE Transaction on Computer Graphics*, 24(1):542–552, 2017.
[9] A. Nakabayashi and T. Itoh. A technique for selection and drawing of scatterplots for multi-dimensional data visualization. In *23rd International Conference on Information Visualisation (IV2019)*, pages 62–67, 2019.
[10] K. Nohno, H.-Y. Wu, K. Watanabe, S. Takahashi, and I. Fujishiro. Spectral-based contractible parallel coordinates. In *18th International Conference on Information Visualisation*, pages 7–12, 2014.
[11] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3):1335–1344, 2012.
[12] L. Shao, A. Mahajan, T. Schreck, and D. J. Lehmann. Interactive regression lens for exploring scatter plots. *Computer Graphics Forum*, 36(3):157–166, 2017.
[13] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.
[14] H. Suematsu, Y. Zheng, T. Itoh, R. Fujimaki, S. Morinaga, and Y. Kawahara. Arrangement of low-dimenional parallel coordinate plots for high-dimensional data visualization. In *17th International Conference on Information Visualisation*, pages 59–65, 2013.
[15] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):759–769, 2020.
[16] A. Watanabe, T. Itoh, M. Kanazaki, and K. Chiba. A scatterplots selection technique for multi-dimensional data visualization combining with parallel coordinate plots. In *21st International Conference on Information Visualisation (IV2017)*, pages 78–83, 2017.
[17] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization*, pages 157–164, 2005.

[18] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2625–2633, 2013.

[19] Z. Zhang, K. T. McDonnel, E. Zadak, and K. Muller. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):289–303, 2015.

[20] Z. Zhang, K. T. McDonnell, and K. Mueller. A network-based interface for the exploration of high-dimensional data spaces. In *IEEE Pacific Visualization Symposium 2012*, pages 17–24, 2012.

[21] Y. Zheng, H. Suematsu, T. Itoh, R. Fujimaki, S. Morinaga, and Y. Kawahara. Scatterplot layout for high-dimensional data visualization. *Journal of Visualization*, 18(1):111–119, 2015.