Brush-Based Ranking For Navigating Within High-Dimensional Datasets

Wolfgang Berger*

VRVis Research Center
Vienna / Austria

Abstract

The analysis of high-dimensional data means a big challenge, as most common visualization techniques do not scale well for displaying a large number of attributes at one time. Therefore, the initial questions arising when analyzing a new dataset typically concern the dimensions themselves in order to assess the relevance of various attributes and to identify clusters of similar (i.e., highly correlated) attributes. After considering this first step, entry-related tasks like detecting outliers or clusters of similar entries can be dealt with more efficiently in a second step. In this paper, we describe an approach which guides the user through a high-dimensional dataset by ranking dimensions and pairs of dimensions according to a large number of statistical summaries. The option to restrict the computations to subsets of the data (e.g., interactively defined by brushing a linked view) and to statistically compare various subsets makes this approach even more powerful and widely applicable, as illustrated by means of a biological dataset.

Keywords: Visual Analytics, Ranking, High Dimensionality, Linking+Brushing

1 Introduction

The steadily rising acquisition, processing, and storage capabilities of current information and communication technologies lead to a growing amount of collected and generated data. For many application domains, this data bears an enormous potential for gaining knowledge and supporting decision-making. Technologies like statistics, datamining and information visualization (InfoVis) address the highly non-trivial issue of extracting useful information from potentially huge datasets in an efficient way.

Statistics have been used for centuries to describe data characteristics in a numerical, easily comparable way. Basic statistical moments like mean, variance or linear correlation are widely used and well-understood and can be calculated in near real-time even for millions of values on today's computers. However, statistics as such is a quite static approach, which hardly involves the user and eventually yields a result without explanation and thus lends itself more to investigating well-defined aspects than exploring new data.

InfoVis on the other hand intends to combine appropriate visual representations of a dataset with means of interaction and thus follows a user-centric approach, which is particularly suitable for exploratory analysis. While being powerful for detecting patterns that could not easily be found by purely mathematical means, visualization alone does not support the user in extracting precise values, as ultimately required in many applications to characterize certain facts and features. Furthermore, most visualization techniques are either inherently limited with respect to the number of simultaneously shown dimensions (like scatterplots), or do not scale well to a large number of dimensions (like parallel coordinates). Therefore analyzing high-dimensional datasets is a big challenge, as users typically do not know, where to look at first and may easily lose the overview.

Comparing the pros and cons of statistics and InfoVis, it turns out that both approaches complement each other very well, which suggests combining both methods in one approach. Especially for the task of analyzing high-dimensional datasets, a reasonable workflow is to base an initial assessment of dimensions and pairs of dimensions on statistical measures, which may eventually guide a subsequent visual and interactive analysis, as described by Seo and Shneiderman by means of their rank-by-feature framework [10, 11].

This paper is structured as follows: After outlining the state-of-the-art in high-dimensional data analysis and briefly introducing the system where this work has been integrated, section 3 describes the extended rank-by-feature framework. In addition to characterizing the whole dataset by means of statistics, this framework explicitly supports multiple subsets (i.e., brushes defined in another view), which may also be subject to frequent changes, and allows for efficient comparisons between them. Section 4 illustrates the usefulness of this technique by analyzing a biological dataset.

2 Related Work

This section summarizes selected approaches for efficiently handling multidimensional data. An emphasis is

*wberger@vrvis.at

put on techniques integrating statistics and visualization.

2.1 Multidimensional Data Analysis

A common way of visually analyzing multi- and high-dimensional datasets is to reduce the amount of dimensions to a level where traditional visualization techniques like scatterplots are applicable. Integrating the user in the reduction process increases the confidence in the result and improves the quality of the visualization compared to fully automatic approaches.

Friedman's and Tukey's *Projection Pursuit* [6] reduces data dimensionality by linearly combining attributes. The user controls the projection process by choosing between several intermediate results in order to get a meaningful final visualization. However, the outcome may be hard to interpret due to the linear combination of probably unrelated attributes. *The Grand Tour* [2], tries to improve this by presenting a set of low-dimensional projections as an animated travel through the dataset, but still requires prior knowledge of the data-characteristics.

In the approach presented Dy and Brodley [5] the system presents different possible dimension subsets based on a user-defined criterion, of which the user selects one considered most useful.

Ankerst et al. [1] introduce a similarity measure, that is used to place dimensions with alike behavior close to each other on the screen. They also provide an approach to solve the arrangement problem for sequential and two-dimensional attribute organization, which uses an ant-system algorithm [4].

Friendly [7] describes a technique for reordering correlation matrices based on a measure that uses the eigenvectors of the matrix and provides a linear arrangement. Furthermore he introduces "corrgrams" for visualizing the results, which is a matrix that is able to reflect the degree of correlation between two dimensions as well as its sign and groups similar attributes.

Yang et al. [16] use hierarchical clustering of similar attributes in order to reduce complexity. After automatically generating the hierarchy, the user can modify it manually in order to improve the resulting visualization.

Guo [8] integrates dimension reduction and sorting as well as data clustering into one framework, the *GeoVISTA studio*. Moreover it allows for user interaction consistently throughout the preprocessing and the visualization process, by letting the user influence the dimension selection and browse the resulting visualization using linking and brushing as well as customizable coloring.

Seo and Shneiderman [11] introduce the rank-by-feature framework, which builds on the *Graphics, Ranking and Interaction for Discovery* (GRID) principles:

- study 1D, study 2D, then find features
- ranking guides insight, statistics confirm.

Histograms and box plots are employed to visualize the distribution of single attributes and 2D scatterplots display all possible pairs of dimensions. Concerning numerical summaries, users may choose between several statistical moments, which are used to rank individual dimensions (1D) or pairs of dimensions (2D). A table displays the resulting ranking along with the exact values of the selected moment. Moreover, a score overview is color-coded and visualizes the 1D case as a list and the 2D case as a triangular scatterplot matrix (SPLOM). This system helps the user finding possibly interesting features in the dataset and improves overview in a high-dimensional dataset.

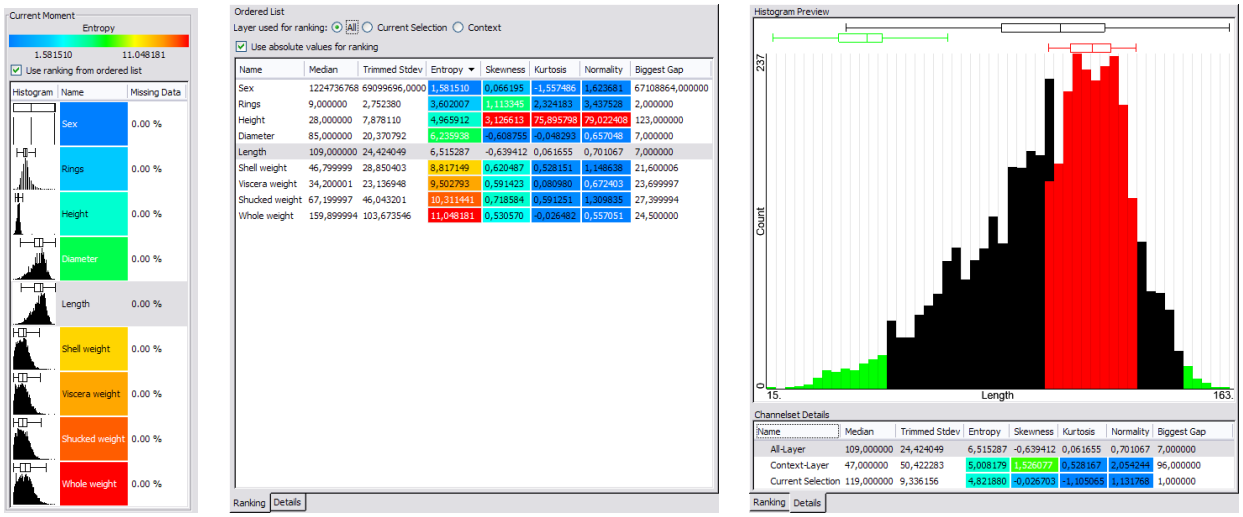
Graph scagnostics as proposed by Wilkinson et al. [14] are an alternative to statistical moments. Initially developed by John and Paul Tukey [13], they characterize two-dimensional point distributions. Dimensionality is reduced by using the results of these calculations and building a feature SPLOM from them. Outliers in this special SPLOM mark unusual 2D scatterplots in the dataset.

Yang et al. [15] introduce a *Value and Relation* display which represents individual dimensions with pixel-based glyphs in order to reveal patterns in the data. These glyphs are positioned in 2D space in a way that relationships between dimensions (such as correlation) are conveyed. The positioning is determined using Multi-dimensional Scaling [9] and places closely related dimensions adjacently to each other. Furthermore a set of interaction tools provides functionality for navigating through the visualization and selecting individual dimensions.

3 Extended Ranking

Although the basic idea for this approach of this paper is inspired by Seo's and Shneiderman's rank-by-feature framework [11], it augments the concept significantly in order to meet the following key requirements:

Query Support – While the original rank-by-feature framework always operates on the whole dataset and is therefore only suited for calculating and presenting global features, the extended approach allows the user to restrict the computation of moments and the resulting ranking to subsets of entries. Our approach has been integrated into a system, where such non-disjunctive subsets are kept in three so-called "data layers", which behave differently concerning their frequency of changes. The "All Entries"-layer represents all data items that are loaded into the system and is not expected to change as long as the underlying dataset does not change. On the other hand, the "Current Selection"-layer consists of entries matching the current query. The user may define and modify such queries via brushing, which typically involves frequent changes. Additional queries are represented by the "Context"-layer which only changes, when whole queries are added, removed or replaced. Keeping the view responsive during potentially time-consuming computations and rapid changes of the considered subset means a significant challenge for the implementation. Moreover, the user is



(a) (b) (c)

Figure 1: The elements of the extended 1D rank-by-feature framework. (a) The score overview showing the value of the current ranking criterion for each dimension in color-coded form along with mini-histograms. (b) The ordered list with a user-defined set of statistical moments. (c) The detail sheet which provides a histogram, a whisker plot and exact values of the statistical moments of the currently selected dimension separately for each layer (data courtesy of Asuncion and Newman [3]).

able to apply layer-based ranking to the color-coded score overview, which helps visually identifying groups of dimensions with similar behavior for the considered subset of the data.

Multiple Feature Calculation and Presentation – In order to be able to check, whether certain dimensions exhibit similar behavior for more than one statistical moment, this approach supports calculating and displaying several numerical summaries simultaneously.

Dimension Selection – Unlike the original rank-by-feature framework, it is possible to assign only a subset of dimensions to the view, in order to be able to handle datasets with several hundred dimensions effectively.

Missing Data Handling – Entries that are known to be missing or invalid must neither influence the outcome of the moment calculations, nor the visualizations.

Large Datasets – The view has to be designed and implemented in a way that it is able to handle large datasets (i.e., up to one million rows and more), which are common in the system it has been implemented for.

3.1 Extended 1D Rank-by-Feature View

According to the additional requirements, the original rank-by-feature framework has been augmented and modified. Figure 1 provides an overview of the organization of its parts.

3.1.1 Multiple Features and Augmented Ranking

A separate control panel allows the user to select, which statistical moments to calculate per dimension and how to arrange them in the ordered list (figure 1 (b)). All offered statistical moments meet the criterion of being computable in a reasonable amount of time even for large datasets. Currently, the available moments are:

- Minimum and maximum.
- Mean and median.
- First and third quartile as well as standard deviation.
- Trimmed mean and trimmed standard deviation: For N entries, the $\lfloor N*0.1 \rfloor$ smallest and $\lfloor N*0.1 \rfloor$ largest values are omitted.
- Skewness, kurtosis and normality: Describe and quantify the deviation from a normal distribution.
- Number of potential outliers: Entries outside $median \pm q_{0.975} * MAD$ (median of absolute deviations) in a set of values that are assumed to follow a standardized normal distribution.
- Entropy: Rises with increasing uniformity of the data distribution.
- Number of unique values.
- Value of the biggest gap.

Finding relationships between different moments of individual dimensions is visually supported. The ordered list applies a transfer function to all numerical summaries that do not depend on an attribute's scale but only rely on the distribution of the data. These color-coded list entries can be easily compared against each other. Furthermore, *every*

(a) (b) (c)

Figure 2: The elements of the extended two dimensional rank-by-feature framework. (a) The score overview with a scatterplot matrix that is color-coded using the transfer function and optionally arranged according to the values of the current ranking criterion. (b) The ordered list, which enumerates all attribute pairs and their corresponding statistical moments. (c) The detail sheet, which provides moment values of the currently selected dimension for each layer (data courtesy of Asuncion and Newman [3]).

column of the ordered list can be used as ranking criterion, there is no discrimination between supplementary information and ranking data as in Seo's and Shneiderman's approach.

Whenever a ranking criterion is set, which also has a transfer function applied, the dimension representations in the adjacent score overview (figure 1 (a)) are also colored accordingly. Moreover, this list contains a mini-histogram and a box-plot for each attribute in order to provide a quick overview concerning the data distribution. In order to group dimensions with similar scores, it is possible to let the system automatically update the ranking of the dimensions based on the current ordering. Another important aspect concerning the score overview is the display of the missing data percentage for each attribute, which allows for roughly assessing the completeness of the underlying dataset.

Especially for the moments which measure the difference from a normal distribution (e.g. skewness, kurtosis, normality), the user might only be interested in the absolute value, not the direction of the aberration. Thus, the possibility of ranking by absolute values is also offered.

3.1.2 Integrating the Concept of Changing Data

At any time, the user is able to switch the data layer the calculations are based on (see top of figure 1 (b)). Furthermore, the user may change the subset defined by the layer itself - e.g., by brushing. This means, that the data and in turns the statistical moments along with the corresponding ranking are subject to change. The concept of asynchronous background computations has been applied in order to keep the view interactive while calculating multiple numerical summaries for potentially millions of entries of a high-dimensional dataset.

The results of the calculations are presented in the visualization as soon as they are available. Whenever the data of a layer changes, the affected measures are re-evaluated immediately and the visualization is updated accordingly, hardly affecting user interactions in the rest of the framework.

By selecting a dimension in either the ordered list or the score overview, the data of all supported layers is visualized in a preview as seen in figure 1 (c). The histogram visualizes the "All Entries"-, "Context"- and "Current Selection"-layers according to a defined drawing order and coloring. Moreover, box plots provide a very condensed overview of the individual data distributions. In the example the current selection (red/dark grey) is drawn on top of the context (green/light gray) which in turn covers the "All Entries" representation (black). Both, the histogram and the box plots, are updated as soon as any of the three layers changes.

The table in the lower part of figure 1 (c) allows for comparing all enabled statistical moments for all considered layers by regarding the currently selected dimension. Thus it complements the ordered list, which displays the moment values of all dimensions and one selected layer.

3.2 Extended 2D Rank-by-Feature View

The extended 2D rank-by-feature view is intended for examining pair-wise relationships for all attributes. Its design closely resembles the 1D counterpart, as shown in figure 2.

Again, a separate control panel offers the possibility to add multiple statistical moments to the ordered list. Currently, the available moments are:

- Pearson's correlation coefficient.
- Spearman's rank correlation coefficient.

While the first is well-known and suited to describing linear relationships, the latter provides more robustness and the ability to detect non-linear dependencies at slightly higher computation costs.

3.2.1 Augmented Score Overview

The dimensions, which have been assigned to the view can be re-ordered in the control panel. This immediately influences the display of the score overview, which is a triangular scatterplot matrix (SPLOM, see figure 2 (a)). The mini-scatterplots are based on the data of all entries and are scaled down to a certain minimal size, if necessary due to screen-space restrictions. However the mini-scatterplot beneath the mouse cursor is always zoomed to its original extents. The zooming is performed smoothly in order to guarantee continuity and avoid change-blindness.

Similar to the 1D case, the score overview lists the precise values of the displayed bivariate moments for all pairs of dimensions. It can be configured to reflect the current ranking of the ordered list (figure 2 (b)). However, mapping a one-dimensional ranking to a two-dimensional arrangement is not straightforward as already mentioned in section 2. Putting the highest ranking pair at the top of the triangular SPLOM turned out to be intuitive. The remaining scatterplots can be arranged according to one of three different mapping strategies:

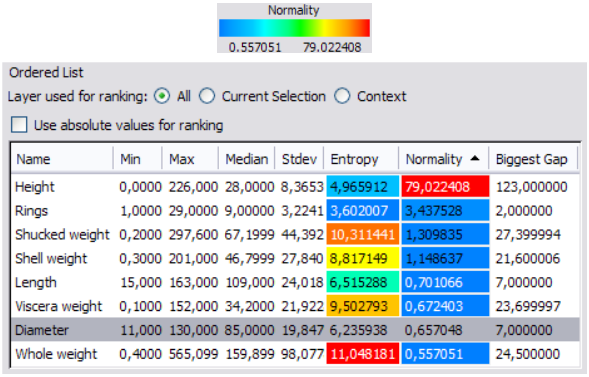
Maximum Ranking – Selects the dimension which produces a SPLOM-row that includes the pair with the highest score of all remaining entries.

Best Average Ranking – Selects the dimension which produces a SPLOM-row, of which the entries generate the highest average ranking. This method is expected to produce a slightly smoother descend from high- to low-ranked attributes for the whole SPLOM.

Diagonal Ranking – Selects the dimension which has the highest ranking when combined with the previously selected dimension. This produces a smooth descend from high- to low-ranked attributes along the diagonal of the SPLOM and tends to form more separate groups of variable combinations.

3.2.2 Layer Support

Multiple, potentially changing layers are supported as for the 1D case. In order to provide a consistent user experience, the same interaction- and processing-concepts have been applied. Furthermore there is a similar detail sheet (figure 2 (c)) available that shows a 2D scatterplot preview and numerical summaries for all layers of one selected attribute pair.



Name	Min	Max	Median	Stdev	Entropy	Normality ▲	Biggest Gap
Height	0,0000	226,000	28,0000	8,3653	4,965912	79,022408	123,000000
Rings	1,0000	29,0000	9,00000	3,2241	3,602007	3,437528	2,000000
Shucked weight	0,2000	297,600	67,1999	44,392	10,311441	1,309835	27,399994
Shell weight	0,3000	201,000	46,7999	27,840	8,817149	1,148637	21,600006
Length	15,000	163,000	109,000	24,018	6,515288	0,701066	7,000000
Viscera weight	0,1000	152,000	34,2000	21,922	9,502793	0,672403	23,699997
Diameter	11,000	130,000	85,0000	19,847	6,235938	0,657048	7,000000
Whole weight	0,4000	565,099	159,899	98,077	11,048181	0,557051	24,500000

Figure 3: The ordered list, which set to rank individual attributes by normality based on the data of the "All Entries"-layer. The according transfer function is depicted above (data courtesy of Asuncion and Newman [3]).

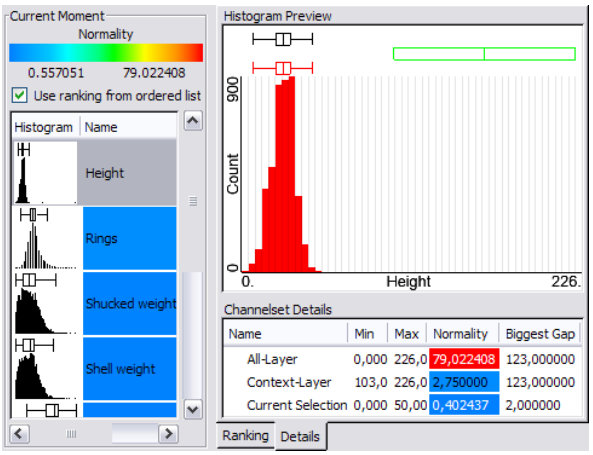


Figure 4: The detail view for the "Height" attribute visualizes the current selection- and context-regions and lists according statistics (data courtesy of Asuncion and Newman [3]).

4 Application

The process of quickly getting an overview over a dataset and extracting first insights using the extended rank-by-feature framework is best described by means of an example. The dataset investigated in this section emanated from a study of abalone populations [3]. In order to determine the age of a single animal, its shell has to be cut through the cone, stained and afterwards the rings have to be counted under the microscope. The gathered attributes for each of the 4,177 examined specimens are:

- Sex (male, female or infant)
- Length (mm)
- Diameter (mm)
- Height (mm)
- Whole weight (gram)
- Shucked weight (gram)
- Viscera weight (gram)

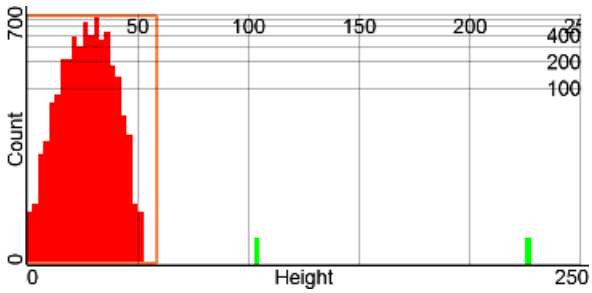


Figure 5: Outliers in the "Height" dimension are separated from the remaining entries by brushing a logarithmically scaled histogram (data courtesy of Asuncion and Newman [3]).

x-Axis	y-Axis	Pearson Corr.	Spearman Corr.
Length	Viscera weight	0,903091	0,952613
Diameter	Shucked weight	0,893385	0,950497
Diameter	Viscera weight	0,899765	0,948338
Length	Shell weight	0,897649	0,947869
Shucked weight	Viscera weight	0,931867	0,947632
Shell weight	Viscera weight	0,907601	0,938085
Height	Shell weight	0,888982	0,921920
Shell weight	Shucked weight	0,882680	0,917741
Height	Whole weight	0,887714	0,916473
Height	Viscera weight	0,865712	0,901123
Diameter	Height	0,906088	0,896334
Height	Length	0,899839	0,888857
Height	Shucked weight	0,836402	0,874434
Rings	Shell weight	0,627817	0,692501
Height	Rings	0,609636	0,658450

Figure 6: The ordered list, which ranks attribute combinations by their absolute Spearman correlation coefficient based on the data of the "All"-layer. The according transfer function is depicted above (data courtesy of Asuncion and Newman [3]).

- Shell weight (gram)
- Number of rings (+1.5 gives the age in years)

As this way of retrieving the age is very time-consuming and tedious, the ultimate goal is to be able to predict it from physical measurements. While the extended rank-by-feature framework is not explicitly designed for classifying data, it is still an excellent tool for acquiring important dimension properties and relationships.

To begin with, all attributes except "Sex" (where a numerical evaluation is not sensible) have been assigned to the extended 1D rank-by-feature view, which computes some basic moments like minimum, maximum, entropy, normality, or biggest gap per default. As clearly visible in the ordered list depicted in figure 3, the dimension "Height" has an outstandingly high normality (colored red according to the transfer function) compared to the other attributes. In addition to this, the value for the biggest gap

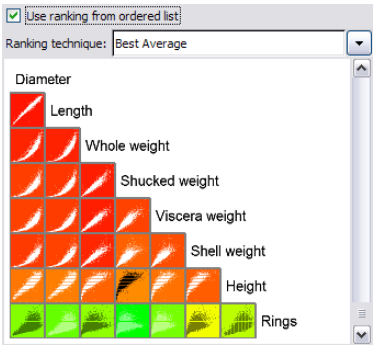


Figure 7: The 2D score overview, which is set to reflect the ranking shown in figure 6 (data courtesy of Asuncion and Newman [3]).

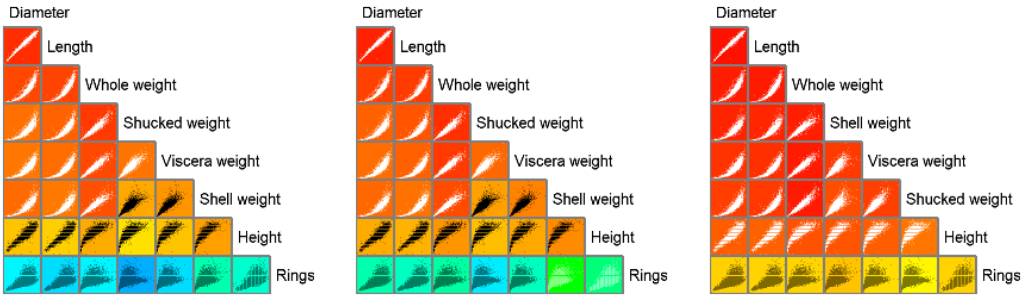
is high in context to the overall value range expressed by the minimum and maximum.

These properties can also visually be verified by examining the mini-histogram of the dimension "Height" in the corresponding entry of the score overview (figure 4 left). The distribution is compressed to the left side, with two low-value histogram bars standing considerably apart. Using the newly introduced layer support, the respective entries can quickly be separated from the remaining data by means of brushing in a linked view (figure 5). The two bars that are considered to be possible outliers are assigned to the "Context"-layer, while the others are set to be in the "Current Selection"-layer as also seen in the histogram in the upper right part of figure 4. The table below lists the statistics of "Height" for all three layers and, as expected, exposes a significantly lower normality.

Further investigation of the suspicious subset reveals that it consists of two entries. The first one describes a male specimen with a height of 103 millimeters and has 10 rings, while the second one refers to a female creature, which is 226 millimeters high and has 8 rings. A quick look at the statistics for "Rings" (figure 3) shows that the median is 9, so that the exceptional values for height obviously do not lead to an exceptional number of rings. Thus these two entries may be considered as outliers and are removed before continuing the investigation, as they might hamper parameterizing a possible classifier.

The reduced dataset is now analyzed with the extended 2D rank-by-feature view in order to find correlating dimensions and in turns identify very few relevant attributes. Again, all attributes besides "Sex" are assigned and the calculation of both, Pearson's and Spearman's correlation are activated for all entries.

As illustrated in figure 6, the Spearman correlation coefficient is generally higher than Pearson's correlation for most pairs. Investigating the mini-scatterplots of the SPLOM shown in figure 7 reveals the reasons for that: Many dimension pairs seem to correlate non-linearly, a property that the Spearman correlation coefficient is more appropriate to detect. Thus this coefficient is used for the next step of the analysis.



(a) (b) (c)

Figure 8: The 2D score overview colored according to the transfer function and the Spearman correlation coefficient for: (a) female, (b) male, (c) infant abalone (data courtesy of Asuncion and Newman [3]).

Figure 7 also demonstrates the mapping of the one-dimensional ranking to the two-dimensional scatterplot matrix. The best scoring attribute pair ("Diameter" vs. "Length") is placed on top followed by the attribute that meets the "Best Average" selection criterion and so on. This reordering makes it easy to identify strongly correlated dimensions, which in this case are "Diameter", "Length", and the different weight measurements. As these attributes behave similarly, it may not be necessary to include all of them in a prediction heuristic. The SPLOM also clearly highlights that there is no single dimension that correlates well with the number of rings. Thus, if the rings (and the age) can be predicted by looking at the available measures at all, it will obviously be necessary to do so by assessing combinations of multiple attributes.

Looking at the mini-scatterplots of the score overview containing the "Rings" attribute reveals another interesting fact: At low ages (and therefore a low number of rings), the point distribution seems to be narrower before spreading out with rising number of rings. Consequently, a separate investigation of the properties of infant specimen is made.

The three different values for "Sex" are brushed sequentially in a separate view and the extended 2D rank-by-feature view is set to use the "Current Selection"-Layer for ranking. While the female and male specimen perform similarly weak concerning the correlation of "Rings" vs. all other dimensions with values < 0.4, the entries representing infant score significantly better with correlation coefficients > 0.7. The different behavior is quickly identified by visually comparing the coloring of the mini-scatterplot backgrounds in the respective bottom rows of the SPLOMs shown in figure 8. This leads to the conclusion that it might be wise to define separate predictors for estimating the age of infant abalone.

To summarize, by using the rank-by-feature views, outliers were easily identified and within a few minutes valuable knowledge has been gained which may significantly speed up the main task, the definition of a prediction heuristic.

5 Conclusions and Future Work

The presented approach is based on the tried and tested [12] rank-by-feature approach and extends it with query support, simultaneous calculation of multiple features as well as design for large datasets. Furthermore it delivers an increased amount of information by always showing mini-histograms (mini-scatterplots) for the assigned dimensions (dimension pairs) and optionally reordering them according to the current ranking. The main contribution is the support for brushes which applies to the visualizations as well as the calculated statistical moments. The possibility of looking for certain features in data subsets allows for more fine-grained investigations and may reveal properties that would have been unnoticed otherwise. The integration of linking makes this functionality even more valuable, as the moments and the ranking are updated in real-time and therefore offer immediate feedback while the user is brushing the dataset, searching for interesting features. This way, the views provide a kind of "statistical characterization" of single brushes, which could be regarded as a way of adding semantical meaning apart from the very definition of the brushes itself.

This elevates the extended rank-by-feature framework to a tool, that can also be used in later stages of the investigation process. When the user has already identified and selected interesting subsets of the data, the support for iterative analysis allows for continuing research based on numerical summaries.

An area that might be subject to further research are special visualizations of statistical moments in the histogram- and scatterplot previews like distribution curves or regression lines. However, when computing and visualizing more than one moment, care has to be taken to prevent cluttering the view. Furthermore the addition of more complex statistical moments (especially in the 2D case) or graph scagnostics [14] (see section 2) could be valuable for users with deeper mathematical knowledge. Finally, explicit support for categorical data, which gets no special treatment at the moment, would make the extended rank-by-feature framework even more useful for datasets like poll-results or census data.

6 Acknowledgments

This work was done at the VRVis Research Center in Vienna, Austria, and was funded by the Austrian research program called Kplus. Thanks go to Helwig Hauser for his supervision and valuable input on this paper.

References

[1] Mihael Ankerst, Stefan Berchtold, and Daniel A. Keim. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, pages 52–60, Washington, DC, USA, 1998. IEEE Computer Society.

[2] Daniel Asimov. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[3] Arthur Asuncion and David J. Newman. UCI Machine Learning Repository (http://www.ics.uci.edu/∼mlearn/mlrepository.html), January 2008.

[4] Marco Dorigo and Luca M. Gambardella. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Transactions on Evolutionary Computation*, 1(1):53–66, April 1997.

[5] Jennifer G. Dy and Carla E. Brodley. Visualization and Interactive Feature Selection for Unsupervised Data. In *KDD '00: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 360–364, New York, NY, USA, 2000. ACM Press.

[6] Jerome H. Friedman and John W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974.

[7] Michael Friendly. Corrgrams: Exploratory Displays for Correlation Matrices. *The American Statistician*, 19:316–325, 2002.

[8] Diansheng Guo. Coordinating Computational and Visual Approaches for Interactive Feature Selection and Multivariate Clustering. *Information Visualization*, 2(4):232–246, 2003.

[9] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. SAGE Publications, 1978.

[10] Jinwook Seo and Ben Shneiderman. A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections. In *INFOVIS '04: Proceedings of the 2004 IEEE Symposium on Information Visualization*, pages 65–72, Washington, DC, USA, 2004. IEEE Computer Society.

[11] Jinwook Seo and Ben Shneiderman. A Rank-By-Feature Framework for Interactive Exploration of Multidimensional Data. *Information Visualization*, 4(2):96–113, 2005.

[12] Jinwook Seo and Ben Shneiderman. Knowledge Discovery in High-Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework. *IEEE TVCG*, 12(3):311–322, 2006.

[13] John W. Tukey and Paul A. Tukey. Computer Graphics and Exploratory Data Analysis: An Introduction. In *Proceedings of the 6th Annual Conference and Exposition: Computer Graphics 85*, pages 773–785, Fairfax, VA, USA, 1985.

[14] Leland Wilkinson, Anushka Anand, and Robert Grossman. High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions. *IEEE TVCG*, 12(6):1363–1372, 2006.

[15] Jing Yang, Anilkumar Patro, Shiping Huang, Nishant Mehta, Matthew O. Ward, and Elke A. Rundensteiner. Value and Relation Display for Interactive Exploration of High Dimensional Datasets. In *INFOVIS '04: Proceedings of the 2004 IEEE Symposium on Information Visualization*, pages 73–80, Washington, DC, USA, 2004. IEEE Computer Society.

[16] Jing Yang, Wei Peng, Matthew O. Ward, and Elke A. Rundensteiner. Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration Of High Dimensional Datasets. In *INFOVIS '03: Proceedings of the 2003 IEEE Symposium on Information Visualization*, pages 105–112, Seattle, WA, USA, 2003. IEEE Computer Society.