Correlation Coordinate Plots: Efficient Layouts for Correlation Tasks

Hoa Nguyen[1(✉)] and Paul Rosen[2]

[1] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, USA
hoanguyen@sci.utah.edu

[2] Department of Computer Science and Engineering, University of South Florida, Tampa, USA

Abstract. Correlation is a powerful measure of relationships assisting in estimating trends and making forecasts. It's use is widespread, being a critical data analysis component of fields including science, engineering, and business. Unfortunately, visualization methods used to identify and estimate correlation are designed to be general, supporting many visualization tasks. Due in large part to their generality, they do not provide the most efficient interface, in terms of speed and accuracy for correlation identifying. To address this shortcoming, we first propose a new correlation *task-specific* visual design called Correlation Coordinate Plots (CCPs). CCPs transform data into a powerful coordinate system for estimating the direction and strength of correlation. To extend the functionality of this approach to multiple attribute datasets, we propose two approaches. The first design is the Snowflake Visualization, a focus+context layout for exploring all pairwise correlations. The second design enhances the CCP by using principal component analysis to project multiple attributes. We validate CCP by applying it to real-world data sets and test its performance in correlation-specific tasks through an extensive user study that showed improvement in both accuracy and speed of correlation identification.

Keywords: Correlation identification · Correlation visualization · Multidimensional data visualization

1 Introduction

Correlation is a powerful metric that provides a predictive relationship between variables used in science, engineering, and business [17,26,32]. A correlation coefficient is a measure of the strength and direction of such a relationship. While correlation is a powerful metric, visual examination is also critical. The many-to-one relationship between data and a correlation coefficient may obscure important features of the data. In Anscombe's Quartet (see Fig. 1) [1], 4 distributions (i.e. the many relationship) have identical correlation coefficients (i.e. the

J. Braz et al. (Eds.): VISIGRAPP 2016, CCIS 693, pp. 264–286, 2017.
DOI: 10.1007/978-3-319-64870-5_13

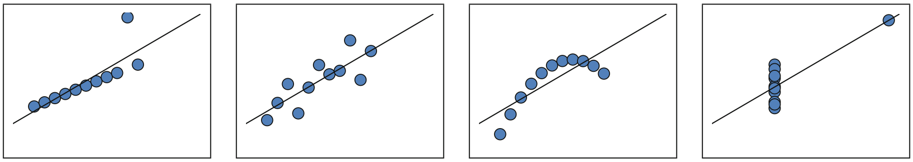


Fig. 1. Anscombe's Quartet [1] shows 4 distributions with an outlier, noise, non-linearity, and non-relationship, respectively, that all have correlation coefficients of 0.816.

one relationship). Visual examination can disambiguate the variations to outliers (case 1), noise (case 2), non-linearity (case 3), and non-relationship (case 4).

Both scatterplots (SCP) [20] and parallel coordinates plots (PCP) [19] are capable of being used to investigate correlation. However, that does not mean one should not infer that these are the *optimal* tools for performing such tasks. In situations where correlation is the most important data feature, these encodings are arguably non-optimal [12,22]. This challenge is exacerbated by the increasing desire to analyze multi-attribute data. A number visualization techniques exist for this analysis [2,4,29], with Scatterplot Matrices (SPLOMs) and PCPs remaining the most popular. SPLOMs simultaneously show all possible combinations of attribute, but the plots become small as the number of combinations grows quadratically. For PCPs, the series of axes grow linearly, but the interface relies heavily upon interaction.

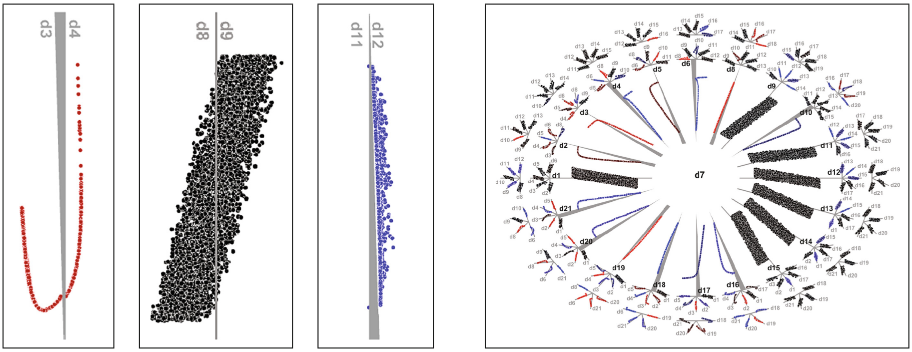
The critical shortcoming to these methods is in their design goal—they are designed as general-purpose tools for performing a wide variety of analytic tasks. No special consideration has been made to any single task, meaning that while they *can be* used to identify correlation, they are *not designed* for it.

With these limitations in mind, we have developed a new, *correlation task-specific* visual design called Correlation Coordinate Plots, or CCPs (see Fig. 2(a–c)). CCPs use design attributes, such as axis shape and a simple, yet effective, point transform to enable quick and accurate determination of correlation direction and strength.

To support multi-attribute analysis we developed 2 different approaches. The first is a focus+context style circular layout for CCPs, called the Snowflake Visualization (see Fig. 2(d)). This visualization represents a compromise where the screen space needed to represent additional attributes grows linearly in the focus and quadratically in the context region. Interaction is still relied upon for full investigation. In the second approach, we have extended the visual metaphors of the CCP to support a single visual interface for multi-attribute analysis by using principal component analysis (PCA) of the data.

To validate the efficacy of our new approaches, use case examples and a user study are used. Our user study had novice and expert subjects perform correlation-related tasks in SCP, PCP, and CCP environments. Our results confirmed that CCP methods outperform SCP and PCP in accuracy and timing.

In summary, the contributions of this paper are:



(a) Positive (b) No Corr (c) Negative (d) Snowflake Visualization

Fig. 2. Correlation Coordinate Plots (CCPs) transform data into a coordinate system better suited to investigating correlation between attributes. (a–c): Example CCPs show positive, no, and negative (or anti-) correlation, respectively. (d): Snowflake Visualization is a focus+context interface that combines CCPs for 1 attribute to all others in the middle (i.e. the focus) and CCPs for all other pairings on the perimeter (i.e. the context).

- a task-specific visualization, the Correlation Coordinate Plot, designed to efficiently identify correlations;
- a circular layout, the Snowflake Visualization, that provides an efficient focus+content style visualization of all pairwise relationships in multi-attribute data;
- a single plot visualization for exploring multi-attribute correlations using PCA; and
- a use case analysis and user study confirming the superior performance of CCP with correlation-related tasks when compared to SCP and PCP.

2 Related Work

2.1 Correlation

Correlation is a metric calculated on data that can be used to model and predict relationships [17,32]. The "quality of relationship" is often measured using a correlation coefficient [6,31], with positive correlation indicating 2 attributes are increasing together, while negative or anti-correlation indicates that 1 attribute increases and the other decreases. There are several correlation coefficient measures, the most common of which is the Pearson Correlation Coefficient (PCC) [25,28]. PCC, $\rho(x,y)$, measures the linear relationship between 2 attributes x and y with means $\bar{x}$ and $\bar{y}$ and standard deviations σ_x and σ_y. It is defined as:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}. \tag{1}$$

As far as correlation in visualization is concerned, there are 2 schools of thought. The first is to show metrics on data, not the data themselves. Examples include Corrgrams [10] and Scagnostics [7,30]. These approaches have the advantages of visual scalability but the potential disadvantages demonstrated by Anscombe's Quartet [1]. An overview of the metrics used in these approaches can be found in [3].

The alternative approach shows all data points. In this category, scatterplots and parallel coordinates have been shown most effective [12,22]. Since our approach follows this paradigm, we compare against these techniques.

2.2 Scatterplot

A Scatterplot (SCP) [5,20] is a simple plot of points used to investigate the relationships between 2 attributes [13]. The patterns of importance in this context are when the data points slope from lower left to upper right, suggesting positive correlation, and sloping from upper left to lower right suggests negative correlation. The direction of correlation (positive or negative) can be confusing to novice users. More importantly, the strength of correlation (high versus low) can at times be difficult to interpret.

For multi-attribute data, a Scatterplot Matrix (SPLOM) [13,18] shows the relationships of all pairs of attributes by organizing a grid of SCPs with each attribute occupying 1 row and 1 column. As the number of attributes increases, the number of plots grows quadratically making it difficult to present all of the data. This problem can be mitigated by approaches such as Corrgrams [10], which display a matrix of correlation glyphs. These glyphs scale well and give the user quick access to summary statistics, but they may hide important data features (e.g. Anscombe's Quartet). In other cases, navigation can be used to search larger spaces [8].

2.3 Parallel Coordinates Plot

Parallel Coordinates Plots (PCPs) [9,19] are another well-known visualization technique for exploring multi-attribute datasets, which display n parallel axes, 1 for each attribute. Data points map to vertices on each parallel axis and connect with line segments. For PCPs, in simple cases, the direction of correlation, though not necessarily intuitive, is easy to identify. Positive correlation appears as a series of parallel lines, while negative correlation appears as crossing lines.

In noisy cases, the ambiguity created by the crossing lines hides patterns but retains outlier visibility [33,34]. This makes correlation direction and strength difficult to interpret. Modifications to PCPs have been proposed by using color, opacity, smooth curves, frequency, density or animation [11,15,16] to partially address this. However, previous studies have shown that PCPs are slower and less accurate than SCPs for correlation tasks [12,22,23].

The advantage of a PCP is that it provides a continuous and comparative view across the axes, and the screen space needed for the visualization scales linearly with the number of attributes. At the same time, PCPs do not show all possible combinations of attribute pairs, requiring significant user interaction for exhaustive exploration. Using 3D parallel coordinates can enable exploration of the many-to-one relationship [21] with the traditional downsides of 3D—perspective effects and occlusion in large data. A PCP matrix [14] is another method that may help overcome this limitation.

3 Correlation Coordinates Plot

The task generality (i.e. the support for many tasks) plays as both an advantage and disadvantage for the SCP and PCP. Either method is capable of being used for correlation tasks, but they are not necessarily the most efficient methods available. This has led us to develop a new visual encoding focused specifically on correlation tasks, called Correlation Coordinate Plots (CCPs). The proposed method is centered on helping users quickly identify the existence, direction, and strength of pairwise correlations. The visual design is motivated by our desire to make the correlation task one of comparison using position along a common baseline, a highly effective visual channel [24].

For clarity in notation, we assume a dataset X contains n attributes and m data points, with X_i indicating a single data attribute of m values and X_{ij} indicating data point j of attribute i.

3.1 Coordinate System

We propose using a correlation coordinate system that differs from the Cartesian coordinate system, so as to highlight how well points adhere to the correlation. The coordinate system can be seen as a 1D parametrization of the data to an underlying model, in this case a line. The vertical position of a data point is the parameterization of the data. The position horizontally is more important, demonstrating the quality of the fit. Therefore, identifying correlation primarily relies on visibility of points to the left and right of the axis.

Transforming the data from a Cartesian domain into the correlation coordinate system is a two step process laid out in Fig. 3, with the top panel showing the positive relationships and the bottom panel demonstrating the negative relationships.

The first step is a scaling operation (Scl) that forces the data into a square region (see Fig. 3 panels 1 & 2). The process begins by normalizing the data to $[-1, 1]$,

$$Scl(X_i) = \frac{X_i - \arg\min_{X_i} X_{ij}}{\arg\max_{X_i} X_{ij} - \arg\min_{X_i} X_{ij}}. \tag{2}$$

The second step is the projection (P_{major} and P_{minor}) operation, which measures the location of the point relative to the positive correlation diagonal (lower left to upper right) or negative correlation diagonal (upper left to lower right). That measure is used to place the points into the CCP (see Fig. 3 panels 3 & 4). The now normalized location of a point i from attributes j and k determines the major (vertical) axis by:

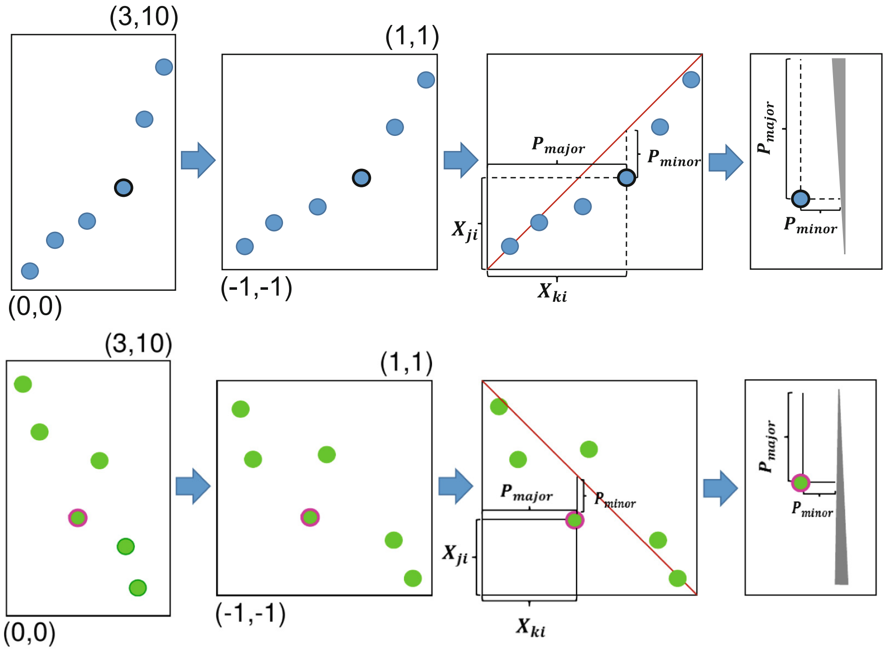


Fig. 3. Conversion to correlation coordinate system for positive (top) and negative (bottom) cases.

$$P_{major}(X_{ji}, X_{ki}) = X_{ki}. \tag{3}$$

The position on the minor axis is:

$$P_{minor}(X_{ji}, X_{ki}) = \begin{cases} \alpha \cdot (X_{ji} - X_{ki}) & \text{positive or no correlation} \\ \alpha \cdot (X_{ji} + X_{ki}) & \text{negative correlation} \end{cases} \tag{4}$$

The variable α is a scalar that effects the spread of data points when plotting. We selected a constant value based upon the width of the CCP.

The plot orientation was initially chosen to be vertical in order to pack many plots side by side on the display. Ultimately, the choice of a vertical plot is somewhat arbitrary and will be relaxed in forthcoming sections. Nevertheless, we present and evaluate our approach based upon the vertical orientation.

3.2 Coordinate Axis

We designed the coordinate axis to serves as a visual indicator of the existence and direction of correlation. For 2 attributes of a dataset, X_i and X_j, PCC is used to indicate positive correlation by $\rho(X_i, X_j) > \epsilon$, negative correlation by $\rho(X_i, X_j) < -\epsilon$, and uncorrelated by all other values. The major coordinate axis is laid out vertically and represented by a triangle whose base is at the top for positive correlation (Fig. 2(a)), the bottom for negative correlation (Fig. 2(c)), and a straight line for uncorrelated (Fig. 2(b)) data.

We have also considered mapping PCC to the width of the axis, where higher values are wider and lower values thinner. Due to the relatively small width of the axis, we decided this mapping was not particularly informative. Instead, to identify the strength of correlation, users should investigate the distribution of data in the correlation coordinate system.

3.3 Coloring Data Points

A number of figures have had their data points colored based upon their PCC value $[\{-1 : blue\}, \{0 : black\}, \{1 : red\}]$. Strictly speaking, this encoding is redundant and not required. However, if colors are interpolated based upon PCC value, they do carry some additional information, and in general, we find them more aesthetically pleasing. Because our focus is on the use of the coordinate axis and coordinate system, our method does not rely on color, and color was *not* used in the user study to be described in Sect. 8.

3.4 Using CCPs for Correlation Identification

Using CCPs for correlation tasks is fairly simple. Depending upon your goal, we suggest:

- First, use the axis to determine if the data is positive, negative, or uncorrelated.
- Next, use the shape of the data points to determine the basic relationship between the attributes (i.e. linear, nonlinear, etc.).
- Finally, the distance of the points from the axis can be used to estimate the strength of correlation, with small distances indicating high correlation, and other conditions such as outliers, noise, etc.

For example, in Fig. 2(c), by checking the axis, a negative correlation can be seen. By observing the closeness of the data points to the axis, a strong linear relationship with small amount of noise. On the other hand in Fig. 2(a), the axis indicates positive correlation. From the shape of the data, it is apparent that a nonlinear relationship exists with weak linear correlation properties.

4 Multi-attribute Visualization

Thus far, our approach can be used to investigate pairwise correlation. Our next goal was to develop an approach for investigating multi-attribute data. We began by looking at SPLOMs which have the advantage of showing all possible combinations of attributes at the cost of the number of plots needed growing at a rate of $O(n^2)$. This may leave little screen space for each individual plot. On the other hand, the number of plots in PCPs grow at a rate of $O(n)$ resulting in more available space for each.

4.1 Parallel CCP

In the PCP spirit, we first applied CCPs to multi-attribute data through a series of equally spaced vertical parallel CCP axes, as seen in Fig. 4. To explore additional combinations of attributes, users can drag an axis to configure the corresponding relationship.

Much like PCP, this approach does not provide immediate access to all attribute pairs, instead relying on user interaction to fully explore the data. As a compromise between the plot size benefits of PCPs and the comprehensiveness of SPLOMs, we developed a new correlation visualization layout, the Snowflake Visualization.

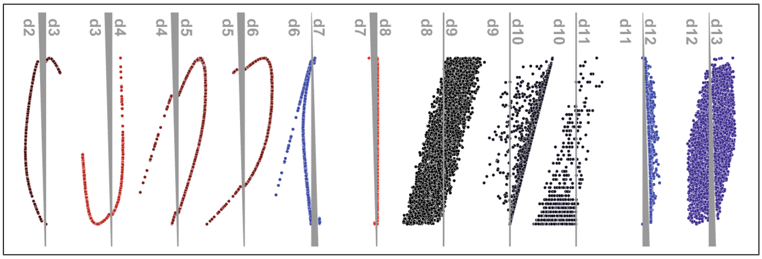


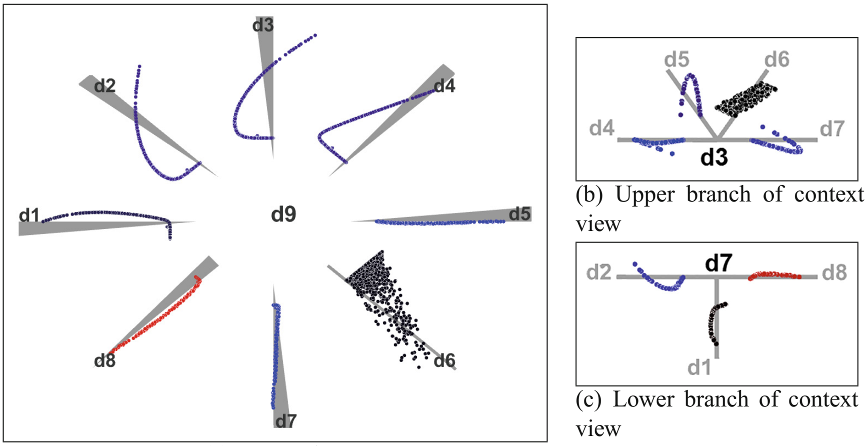
Fig. 4. Parallel CCP for 10 attributes data allow full exploration of the data, but, like PCP, it relies on heavy user interaction.

4.2 Snowflake Visualization

We focused on a radial based design due to their efficient use of space for multiple attribute visualizations [27]. As such, we have developed the Snowflake Visualization, which is constructed of a focus+context views.

Focus View. The focus view (Fig. 5(a)) enables investigating the correlation of 1 attribute to all other attributes. Given n attributes, there are $(n-1)$ pairs laid out around the center of the circle with equal angular spacing. By default, the final attribute of data is the initial focus attribute. Attributes are sorted by ID but can be reordered with other sorting methods. The inner radius (the start of the CCP axes) is chosen such that none of the data points between CCPs will overlap. The outer radius (the end of the CCP axes) is adjustable as to give more or less space to the context views.

Context View. Given the attributes covered by the focus view, we designed the context view to give complete coverage of the remaining attribute pairs. These context views (Figs. 5(b) and (c)) are attached to the branches of the focus view. The objective is to prevent pairs of attributes from being repeated. This is done by organizing the pairings based on parity of n.



(a) Focus view

(b) Upper branch of context view

(c) Lower branch of context view

Fig. 5. A focus view (a) and multiple context views (b–c) for Snowflake Visualization.

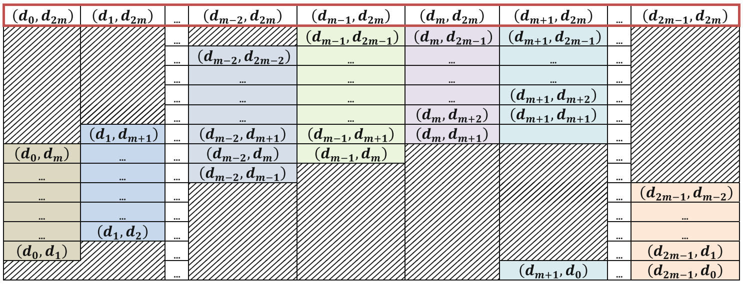


Fig. 6. Branch attribute pairing matrix for n attributes, when n is odd, and the focus attribute is d_{2m}. Each row and column represent 1 data attribute. Pairings are found by selecting a column for the attribute and pairing with highlight attributes.

When n is odd ($m = (n-1)/2$), the organization, shown in Fig. 6, contains all pairwise correlations in data. Each pair (d_i, d_j), where $i = 0, 1, \ldots, 2m-1$ and $j = i+1, i+2, \ldots, 2m$, presents correlation between 2 attributes d_i and d_j. The red box in Fig. 6 contains all the attribute pairs that are presented in focus view, pairing the last attribute d_{2m} and all other attributes $(d_0, \ldots, d_{2m-1})$.

The context view has two groups—the upper branches and lower branches. In the upper branches, where $i = 0, \ldots, m-1$, the i^{th} branch presents correlations between attribute d_i and m other attributes that are $(d_{i+1}, d_{i+2}, \ldots, d_{i+m+1})$. There are m pairwise correlations in each upper branch. We can see one upper branch in the Fig. 5(b) that has 4 pairwise correlations when the number of attributes n is 9 and m is 4.

In the lower branches, where $i = m, m+1, \ldots, 2m-1$, the i^{th} branch presents correlation between attribute d_i and other attributes as shown in the

i^{th} column in Fig. 6. There are $(m-1)$ attribute pairs in each lower branch. Figure 5(c) shows one lower branch that presents 3 pairwise correlations when number of attributes $n=9$ and $m=4$.

The organization is similar when n is even ($m = n/2$). The focus view presents the correlations between last attribute d_{2m-1} and all other attributes in data $(d_0, d_1, \ldots, d_{2m-2})$. The context view has only a single branch type that has $m-1$ attributes pairs in each upper or lower branch.

Detail View and Interaction. Typically a single large CCP detail view is also included with the Snowflake Visualization (a similar practice to SPLOMs). A few interactions are included with the Snowflake Visualization. These include:

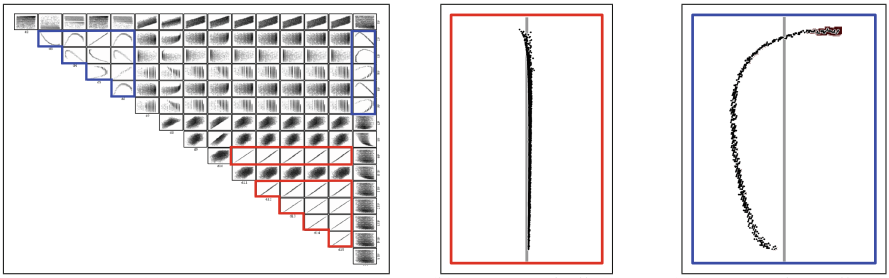
- *Click-to-swap*: When the user clicks an attribute, it becomes the focus attribute. After swapping, outer attributes are reordered based upon a sorting criteria (i.e. by attribute ID).
- *Over-to-detail*: As the mouse moves over a plot, the detail view is updated to that pairing.

5 Multiway Attribute Correlations

Pairwise correlations are frequently important to understanding data. However, as the number of attributes increases, the desire to explore relationships of multiple attributes simultaneously increases as well. The Snowflake Visualization partially addressed the need by presenting many pairwise relationships simultaneously. Comparing 3 or more attributes requires looking at an exponentially increasing number of plots and mentally fusing the distributions. We can extend CCP design for presenting certain types of multi-attribute relationships.

To do this, we slightly modify visual metaphors of the CCP. First of all, we remove the positive/negative metaphor encoded via the axis. This is because multi-attribute relationships tend to not have a directional measure, only magnitude. Now, the parameterization model can be relaxed to any invertible function, $[s,t] = g(\bar{x})$. The vertical axis still represents a 1D parameterization of the data, s. The horizontal axis can now represent a secondary model parameterization, t. Finally, we represent information lost in this encoding via a series of partially transparent boxes, one per data point, that form a "haze" surrounding the data points. The size of the boxes found using the residual, $r = ||\bar{x} - g^{-1}(s,t)||$.

For our experiments we have used Principal Component Analysis (PCA) to parameterize the data. This could be replaced with any other model that fits our functional definition. Using PCA, we set $g(\bar{x})$ equal to the magnitude of the first two principal components of the data, and the size of the box is set to the residual. Figure 7 shows 2 examples. The SPLOM on the left (Fig. 7(a)) shows all of the attributes of the dataset. Two subsets have been selected in red and blue. The red subset are attributes that all appear pairwise linear. When we use the many-attribute CCP (Fig. 7(b)), we can see that all of the attributes are linear with respect to one another. On the other hand, the blue attributes appear nonlinear. When visualized with the many-attribute CCP (Fig. 7(c)), we can see a relatively simple nonlinear 2D pattern within the data.



(a) SPLOM (b) CCP of linear feature (c) CCP of nonlinear feature

Fig. 7. CCP for multiple attributes using PCA. (b) The attributes in red are a linear feature. (c) The nonlinear feature in blue is 2D, with the residual visible in the red haze. (Color figure online)

6 Implementation

Algorithms 1, 2 and 3 contain pseudocode for the CCP and Snowflake Visualization. We have also included a sample visualization tool[1] that can be built in Processing.

Algorithm 1 is pseudocode to draw the CCP of two input data attributes with M items as we described in Sect. 3. Drawing the Snowflake Visualization is presented in 2 parts. The focus view, based on attribute j, can be draw using Algorithm 2 by drawing a series of CCPs plots around the center, with equal angular spacing. Algorithm 3 presents the method to draw context view of Snowflake Visualization, based on the parity of n.

Algorithm 1. Draw Correlation Coordinate Plot.

// Draw axis
if $PCC(X, Y) > \epsilon$ **then**
  $drawAxis$(upper-triangle-axis)
else if $PCC(X, Y) < \epsilon$ **then**
  $drawAxis$(lower-triangle-axis)
else
  $drawAxis$(straight-line-axis)
end if

// Draw items
for $i = 1 : M$ **do**
  $[x_n, y_n] := normalize(X_i, Y_i)$
  $p_{major} := x_n$
  if $PCC(n-1, i) > \epsilon$ **then**
    $p_{minor} := \frac{y_n - x_n}{2.0f}$
  else
    $p_{minor} := \frac{y_n + x_n}{2.0f}$
  end if
  $drawPoint(p_{major}, p_{minor})$
end for

[1] CCPs: https://github.com/hoa84/CCPs_SnowflakeViz.

Algorithm 2. Draw Focus View of Snowflake.

```
// Draw attributes before focus j
for i = 1 : j − 1 do
    setPosition(cen, rad, (i−1) · 360°/(n−1))
    drawCCP(A_j, A_i)
end for
// Draw attributes after focus j
for i = j + 1 : N do
    setPosition(cen, rad, i · 360°/(n−1))
    drawCCP(A_j, A_i)
end for
```

Algorithm 3. Draw Context View of Snowflake.

```
// Parity bit for even vs. odd n
even = (n is even) ? 1 : 0

// Draw attributes after focus j
m = ⌊n/2⌋

// Loop ranges
range_0 := m − even
range_1 := 2m − 1 − even

// Angular separation for plots
b_0 = 180°/(m − 1 − even)
b_1 = 180°/(m − 2)
for i = 0 : range_0 do
    for j = 0 to m − 2 do
        setPosition(cen_i, ang_i + b_0 ∗ j)
        drawCCP(A_i, A_{i+j+1})
    end for
end for
for i = range_0 + 1 : range_1 do
    for j=0 to i-m+1-even do
        setPosition(cen_i, ang_i + b_1 ∗ (2m + j − i − 2))
        drawCCP(A_i, A_j)
    end for
    for j=i+1 to 2m-2 do
        setPosition(cen_i, ang_i + b_1 ∗ (j − i − 1))
        drawCCP(A_i, A_j)
    end for
end for
```

7 Usage Examples

We applied three visualization methods, including the Snowflake Visualization, SPLOM, and PCP, to three publicly available datasets including Boston house price data[2], Pollen data[3], and Hurricane Isabel data[4].

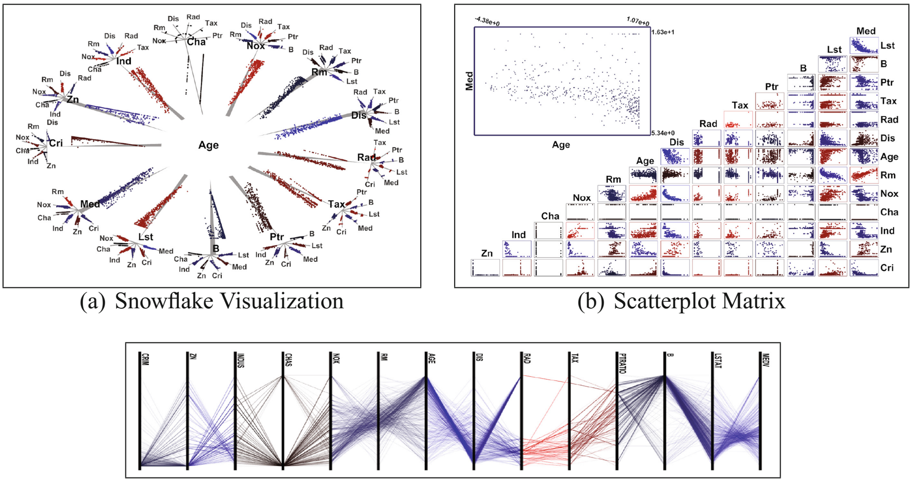
7.1 Boston House Price

Boston housing data (see Fig. 8) is multivariate dataset containing 506 items across 14 attributes. The data contains several variables that try to explain variation in home values in the Boston area.

When comparing this dataset in a Snowflake Visualization and SPLOM, there are a number of features observable in both visualizations. For example, in both visualizations the Age/Rad pairing is fairly clearly a case for segmentation into two data groups. In the SPLOM, it will likely take longer.

A big advantage in Snowflake Visualization is that it makes way for exploiting additional visual channels. Take the Age/Ind pairing. In all visualization

[2] http://lib.stat.cmu.edu/datasets/boston.
[3] http://lib.stat.cmu.edu/datasets/pollen.data.
[4] http://vis.computer.org/vis2004contest/.



(a) Snowflake Visualization

(b) Scatterplot Matrix

(c) Parallel Coordinates

Fig. 8. Visualizations for Boston House data.

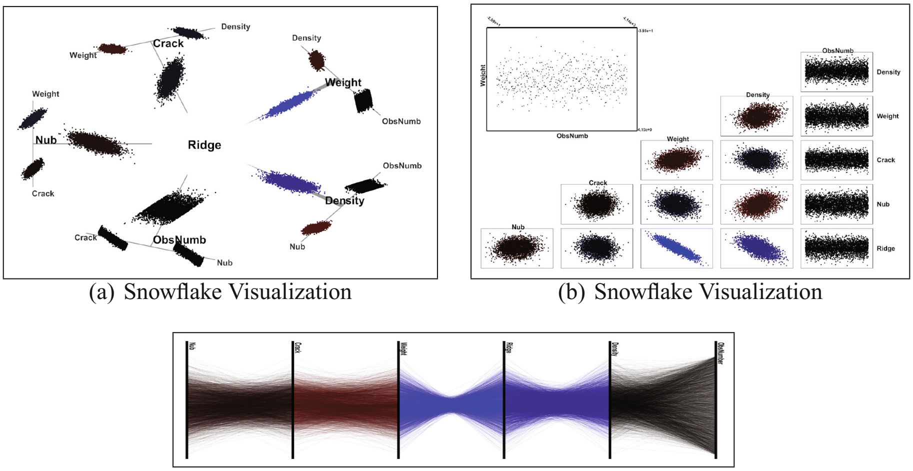
approaches, coloring scheme we have used makes it fairly easy to see that there is a strong positive correlation. However, without the coloring that might not be the case. If color had been used for some other purpose, classification for example, suddenly we lose the ability in SPLOMs to quickly determine correlation, while observing classification. Since CCPs do not rely on color to communicate correlation, we can encode other information in the color channel without significant loss of correlation information.

7.2 Pollen Data

The pollen data (Fig. 9) contains 3848 items each with 6 attributes. This dataset summarizes geometric features of pollen grains.

The nature of the data makes it difficult to use the PCP due to overdraw. Take the Ridge/Weight and Ridge/Density. Even though we can be fairly certainly that Ridge/ Weight is more negatively correlated than Ridge/Density, any other detail is lost. We are unable to determine if it is due to outliers, nonlinearity, noise, etc. Techniques such as clustering, density, histogram PCPs can be used to further improve the representations. However, for correlation strength tasks, these approaches are not particularly beneficial.

For the Snowflake Visualization this data proves little trouble. When Ridge is selected as the focus parameter, Density and Weight can be compared in detail. The thinner spread of Ridge/Weight indicates a stronger linear relationship compared to Ridge/Density. In addition, the details available in the view confirm that any weakness in the correlation is due to noise.



(a) Snowflake Visualization (b) Snowflake Visualization

(c) Parallel Coordinates

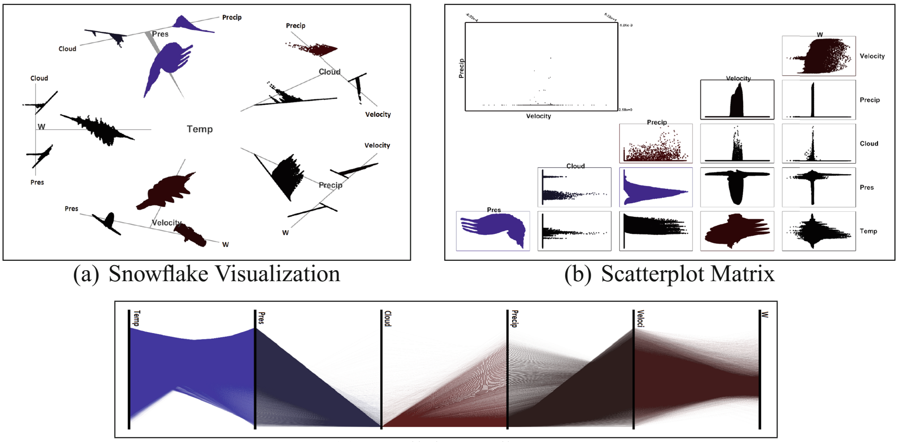
Fig. 9. Visualizations for pollen data.

7.3 Hurricane Data

Hurricane Isabel (Fig. 10) data is provided as part of the IEEE Visualization 2004 contest. This dataset contains a variety of simulated variables related to Hurricane Isabel, a major Atlantic storm that occurred in September of 2003. Isabel data set consists of 48 timesteps, each containing measurements of 11 attributes with a spatial resolution of $500 \times 500 \times 100$. We also only show 7 of the more "interesting" attributes due to space considerations. Of the original data 25 million data items, we only use 10 million because approximately 15 million data items contain at least 1 invalid *NaN* field.

With 10 million data items in the Hurricane data, the overdraw in the PCP makes it hard to understand any relationships in the data. For example, the relationship Temp/Pres shows only the bowtie shape, losing the individual data patterns. In many ways, SCPs do a better job than PCPs. The Temp/Press relationship is visible with the SCP. However, clear interpretation is difficult, since as Temp increases, Press first decreases, then increases, and finally decreases.

Our approach presents these relationships more clearly. The direction and strength of relationship between Temp and Pres can be identified in Snowflake Visualization. The lower triangle shape of axis identifies the negative relationship. Additionally, the data points distribution, mostly being of similar distance to the axis with a few spread out, enables identifying that this relationship is not strongly negative and nonlinear.



(a) Snowflake Visualization (b) Scatterplot Matrix

(c) Parallel Coordinates

Fig. 10. Visualization techniques for Hurricane data.

8 User Study on Identifying Correlation

To further evaluate our visualization methods, we conducted a user study comparing CCP with SCP and PCP. In this study, we performed 3 experiments that ask subjects to perform correlation related tasks.

We invited 25 participants to take part in our study, 9 female and 16 male, all graduate students from a variety of science and engineering fields. Their ages range from 23 to 35 years old. We asked the subjects to self-report their level of familiarity with visualization—3 reported themselves as experts; 9 reported themselves as familiar; and 13 reported themselves as not familiar.

In each experiment, subjects started with a short set of slides and/or video to introduce the necessary background. Subjects were then given practice questions where, after answering, the correct answers were provided. They would then perform the experimental tasks. For each test, the subjects' answers and response times were recorded. Following the experiment, subjects completed a short survey. In total, the study lasted less than one hour, including training and testing. For all visualizations, gray color was used for axes and labels, black color was used to present data items.

The software for the user study was built using C++ and Qt, and run on a MacBook pro with a 2.5 GHz Intel Core i5, 4 GB RAM, and 512 MB Intel HD Graphics 4000. The study used a particle physics dataset containing 41 output attributes and 4000 data items per attribute. The data represents a parameter space search of 25 input attributes generated by a series of tools that simulate the theoretical physical properties of subatomic particles under the Supersymmetic extension of the Standard Model of particle physics.

The independent and dependent variables used in each experiment can be found in Table 1. We used a mixed experimental design using t-testing to

Table 1. Variables used to test hypotheses.

Independent variables	Potential values
Data [**H1** | **H2** | **H3** | **H4**]	2 random attributes from 41 attribute data
Data [**H5** | **H6** | **H7**]	10 or 21 attributes from 41 attribute data
Plot [**H1** | **H2** | **H3** | **H4**]	SCP/PCP/CCP
Plot [**H5** | **H6** | **H7**]	SPLOM/PCP/Snowflake Visualization
Question [**H1** | **H2**]	How are the 2 attributes correlated?
Question [**H3** | **H4**]	What is the type of correlation?
Question [**H5**]	How are the 2 attributes correlated?
Question [**H6**]	How many attributes are correlated to i?
Question [**H7**]	Which attributes are correlated to i?
Dependent variables	Potential values
Answer [**H1** | **H2** | **H5**]	High Positive Correlation
	Low Positive Correlation
	No Correlation
	Low Negative Correlation
	High Negative Correlation
Answer [**H3** | **H4**]	Nonlinear Correlation
	Linear Correlation
	No Correlation
Answer [**H6**]	Number of attribute
Answer [**H7**]	List of attribute
Response Time [all **H**]	Time recorded automatically

calculate t-value, p-value, mean difference, and 95% confidence interval to confirm our hypotheses. Only mean value and p-value are reported, but other data can be provided upon request.

8.1 Exp 1: Speed/Accuracy in Pairwise Correlation

When looking at SCP & PCP, 2 challenges persist. First, it can be confusing to determine positive versus negative correlations. Granted, for experts this is a trivial task, but for others, it can be confusing. In many ways the identification of correlation direction is easier with PCP than SCP—parallel lines positive and crossing lines negative. Second, there is some ambiguity when trying to identify the strength of correlation between 2 attributes. Ambiguity is a much larger problem for PCP. When the relationship is noisy or nonlinear, overlapping lines quickly obscure detail.

When comparing CCP with these other methods, CCP: (1) provides simple visual cues making identification of the direction of correlation fairly trivial; (2)

and reduces (not eliminates) the ambiguity by concentrating on correlation in the formulation of the coordinate system.

Given these factors, we developed 2 hypotheses as follows:

H1 | **H2:** *Using a Correlation Coordinates Plot will enable more accurate and faster identification in direction and strength of correlation between 2 attributes than a [H1: Scatterplot | H2: Parallel Coordinates Plot].*

Method. The experiment is summarized in Table 1 (**H1** & **H2**). For a block of trials, we showed a participant a plot between 2 random attributes using either the SCP, PCP, or CCP method and asked a forced choice question. Subject accuracy and time were measured.

At the start of the experiment, participants were given an introduction to correlation, instructions on finding correlation in SCP, PCP, and CCP, and 6 training questions. Participants were then given 21 experimental questions (7 for each plot type, interleaved order).

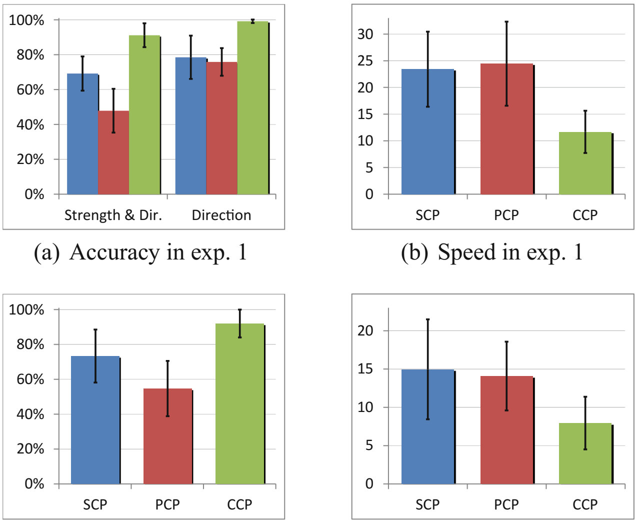
Results and Discussion. The results of both the measured speed and accuracy of our experiments are shown in Fig. 11(a) and (b).

The results from Fig. 11(a) shows that when comparing accuracy, CCP showed improvement over SCP on average 91% compared to 69%, with statistical significance ($p = 0.001$). We also looked at subjects performance in just identifying the direction of correlation, where CCP had an accuracy of 99% compared to 79% for SCP, though not quite with statistical significance ($p = 0.06$). The response times (Fig. 11) showed similar results with CCP responses averaging 11.71 s compared to 23.4 s for SCP ($p = 0.001$). Given that in our experiments CCP outperformed SCP in both speed and accuracy, we consider **H1** confirmed.

A similar analysis shows that the accuracy CCP was 91% compared to 48% for PCP ($p = 0.001$). For identifying type only, CCP had an accuracy of 99% compared to 76% for PCP, though not with statistical significance ($p = 0.09$). The response times (Fig. 11) showed a similar result with CCP coming in on average 11.71 s compared to 24.5 s for PCP ($p < 0.001$). Given that CCP outperformed PCP in speed and accuracy, we consider hypothesis **H2** confirmed.

Although not explicitly selected as a hypothesis, we are also able to compare the performance of SCP and PCP. The results showed that SCP had a higher overall accuracy on average, 70%, compared to 58% for PCP ($p = 0.048$). However, when looking at type accuracy only, SCP had no statistical significance in average accuracy of 76% compared to 85% for PCP ($p = 0.25$). The results showed no statistical significance in average response times of SCP and PCP of 24.54 s and 25.78 s ($p = 0.774$), respectively. This result aligns with prior work [12,22,23].

The results of Exp. 1 confirmed the hypotheses **H1** and **H2**, indicating that using CCP subjects can identify correlation in less time and with higher accuracy compared to SCP and PCP. In our informal discussions with subjects after the experiment, they indicated that the shape of the axis and the distribution of



(a) Accuracy in exp. 1
(b) Speed in exp. 1
(c) Accuracy in exp. 2
(d) Speed in exp. 2

Fig. 11. Results of exp. 1 and exp. 2 show CCP (green, col. 3) outperforming SCP (blue, col. 1) and PCP (red, col. 2) in speed (sec) and accuracy (%). In all figures, error bars indicate standard deviation. (Color figure online)

points in CCP greatly assisted their comprehension of the correlation. Subjects complained that both the SCP and, in particular, the PCP were more difficult to distinguish positive and negative correlation in scenarios with low correlation. However, they found using CCP enabled them to easily recognize both the direction and strength.

8.2 Exp. 2: Differentiating Linear, Nonlinear, and Uncorrelated Relationships

Identifying nonlinear relationships between attributes can also be an important task. When comparing CCP with other methods, CCP provides simple visual cues making identification of correlation direction easier. Beyond that, CCP and SCP give similar visual cues (i.e. the tasks performed are basically the same) for the shape of the relationship, linear or nonlinear. This motivates our next hypothesis:

H3: *Using a Correlation Coordinates Plot and a Scatterplot will result in similar accuracy and speed for identification of linear, nonlinear, and uncorrelated relationships in 2 attributes.*

For PCP, identifying these relationships is far more challenging. The overdraw ambiguity that plagues linear correlations becomes significantly worse as even more lines overlap each other in nonlinear cases. This will slow and confuse users. This leads to our next hypothesis:

H4: *Using a Correlation Coordinates Plot will result in more accurate and faster identification of linear, nonlinear, uncorrelated relationships in 2 attributes than a Parallel Coordinates Plot.*

Method. The experiment is summarized in Table 1 (**H3** & **H4**). At the start of the experiment, participants were given instructions on linear and nonlinear correlation. Participants were then given 3 training questions followed by 9 experimental questions (3 for each plot type, interleaving order). For each question, participants saw a plot from 2 random attributes and were asked a forced choice question. Subject accuracy and time were measured.

Results and Discussion. The results of the measured speed and accuracy of our experiments are shown in Fig. 11(c) and (d), with all differences showing statistical significance ($p < 0.005$). The results of our experiment showed that CCP outperformed SCP. Our hypothesis **H3** however had predicted that the performance of CCP and SCP would be identical. This leads us to reject **H3**. In our discussions with subjects after the experiment, they indicated that the shape of axis and the distribution of points in SCP was more difficult to distinguish and that CCP assisted their comprehension of these specific types of correlation.

Due to CCP substantially outperforming PCP in both speed and accuracy, we consider hypothesis **H4** confirmed. As anticipated, participants complained that the overdraw problems made it difficult differentiate linear vs. nonlinear correlations in PCP.

8.3 Exp 3: Accuracy/Speed in Multi-attribute Datasets

The Snowflake Visualization was designed specifically for the task of quickly and accurately exploring pairwise correlations in multi-attribute data as compared with SPLOMs and PCP. As the number of attributes increases each SCP within a SPLOM becomes quite small and the number of plots becomes overwhelming. For PCP, as the number of attributes increases, the interaction required for many tasks puts increased pressure on the user to explore for features of interest. With these factors in mind, we developed 3 hypotheses:

H5: *Using a Snowflake Visualization will enable more accurate and faster identification of correlation between 2 attributes in multi-attribute data than a Scatterplot Matrix or Parallel Coordinates Plot.*

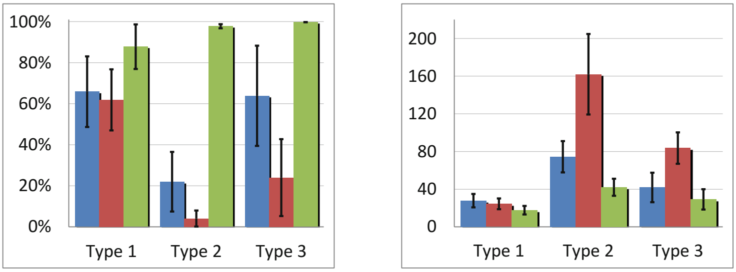
H6: *Using a Snowflake Visualization will enable more accurate and faster identification of how many attributes are correlated with a chosen attribute in multi-attribute data than a Scatterplot Matrix or Parallel Coordinates Plot.*

H7: *Using a Snowflake Visualization will enable more accurate and faster identification of which attributes are correlated with a chosen attribute in multi-attribute data than a Scatterplot Matrix or Parallel Coordinates Plot.*

Method. The experiment is outlined in Table 1 (**H5-H7**). Each participant was given an introduction and demo video for each visualization method and completed 12 sample questions using data unrelated to experimental trials. Then, each performed 21 experimental questions interleaving first between visualization types, then question types.

Results and Discussion

Identification of a Pairwise Correlation in Multi-attribute Data. The results of measured speed and accuracy in Fig. 12 (Type 1) show the Snowflake Visualization improved accuracy and speed over SPLOMs and PCPs with statistical significance (all $p < 0.05$). The average accuracy for Snowflake Visualization was 89% compared to 67% for SPLOM and 64% for PCP. The response times (Fig. 12) for Snowflake Visualization came in at an average of 19.9 s compared to 31.3 s for SPLOM and 26.1 s for PCP. Given that Snowflake Visualization outperformed SPLOM and PCP in speed and accuracy, we consider hypothesis **H5** confirmed.



(a) Accuracy (%) (b) Response Time (sec.)

Fig. 12. Exp. 3 results show CCP (green, col. 3) outperformed SCP (blue, col. 1) and PCP (red, col. 2). (Color figure online)

Finding the Number of Correlated Attributes in Data. Again, the results of the experiments showed that the Snowflake Visualization improved accuracy and speed over SPLOMs and PCPs (see Fig. 12, Type 2) with statistical significance ($p < 0.05$). Therefore, we consider hypothesis **H6** confirmed.

Finding which Attributes are Correlated in Multi-attribute Data. The results of this final test also showed improved accuracy and speed over SPLOMs and PCPs (see Fig. 12, Type 3) with statistical significance ($p < 0.05$), leading us to also consider hypothesis **H7** confirmed.

The Snowflake Visualization's focus+context style greatly assisted subjects interactions and comprehension when working through multiple pairwise correlation questions. The participants complained that small SCPs made the SPLOM difficult to use, due to inability to see individual plots and difficulty tracking rows or columns of plots. Using PCP, participants complained that the number of dragging operations required to explore multiple correlations made it very difficult for them.

9 Discussion

User Study Task Selection. Selecting realistic tasks for a user study is a challenging problem when users are unfamiliar with the data and potentially visualization altogether. We have selected a number of simple tasks, which are building blocks for more complicated data analysis tasks that are commonly performed. The overall out performance of the CCP over SCP and PCP stands as evidence of its superiority, which should translate to more complex tasks.

Abstraction Selection. SCP and PCP have served a straw man role in our evaluation. There are any number of modifications that could be applied to either technique to better inform the user about correlation. However, since there is no single de facto standard, we did not want our evaluation to be clouded by questions of abstraction selection in SCP or PCP. Therefore, we stuck to the basic formulations of each approach. We hope this paper spurs the community to dig deeper into this subject and generate a more extensive evaluation of approaches, such as those of Harrison [12] and Kay [22].

Very High Attribute Count Data. For data with large numbers of attributes, we believe that approaches to extract the natural dimensionality of data, such as PCA, in combination with techniques such as CCP, will be critical in analysis. For all practical purposes, beyond 30 or 40 attributes, our approach is no longer viable. However, this is a similar limitation to SPLOMs and PCPs. We consider higher-dimensional cases to still be an open problem.

10 Conclusion

Correlation Coordinate Plots have been developed with the specific task of correlation identification. They have distinct advantages when compared to general task visualizations such as SCP and PCP. The advantages, as confirmed by our user study and real-world datasets, include:

- providing simple visual cues that make identification of the existence and direction of correlation fairly trivial;
- improving estimation of correlation strength by focusing the coordinate system on model fit; and
- improving identification of linear, nonlinear, and uncorrelated data by reducing ambiguity in the visualization.

In addition, the Snowflake Visualization showed significant performance improvements over SPLOMs and PCPs. The Snowflake Visualization is an efficient focus+context style layout representing a fair compromise between space efficient design, comprehensive visualization, and reduced user interaction for showing all pairwise correlations in multi-attribute data.

In conclusion, we believe that the CCP and Snowflake Visualization represent complementary approaches to existing techniques, replacing existing approaches only where correlation is the major feature of focus in data. We believe that more of these task specific approaches are on the horizon and will provide data analysts better, faster access to relevant information in their data.

Acknowledgements. We would like to thank our reviewers and colleagues who gave us valuable feedback on our approach. We would also like to thank our funding agents, NSF CIF21 DIBBs (ACI-1443046), NSF Core Program (IIS-1513616), Lawrence Livermore National Laboratory, and Pacific Northwest National Laboratory Analysis in Motion (AIM) Initiative.

References

1. Anscombe, F.J.: Graphs in statistical analysis. In: American Statistical Association, pp. 17–21 (1973)
2. Aris, A., Shneiderman, B.: Designing semantic substrates for visual network exploration. InfoVis **6**, 281–300 (2007)
3. Bertini, E., Tatu, A., Keim, D.: Quality metrics in high-dimensional data visualization: an overview and systematization. IEEE Trans. Vis. Comp. Graph. **17**(12), 2203–2212 (2011)
4. Bezerianos, A., Chevalier, F., Dragicevic, P., Elmqvist, N., Fekete, J.D.: Graphdice: a system for exploring multivariate social networks. Comput. Graph. Forum **29**(3), 863–872 (2010)
5. Buering, T., Gerken, J., Reiterer, H.: User interaction with scatterplots on small screens - a comparative evaluation of geometric-semantic zoom and fisheye distortion. IEEE Trans. Vis. Comp. Graph. **12**(5), 829–836 (2006)
6. Chen, Y.A., Almeida, J.S., Richards, A.J., Muller, P., Carroll, R.J., Rohrer, B.: A nonparametric approach to detect nonlinear correlation in gene expression. J. Comput. Stat. Graph. **19**(3), 552–568 (2010)
7. Dang, T.N., Wilkinson, L.: Transforming scagnostics to reveal hidden features. IEEE Trans. Vis. Comp. Graph. **20**(12), 1624–1632 (2014)
8. Elmqvist, N., Dragicevic, P., Fekete, J.D.: Rolling the dice: multidimensional visual exploration using scatterplot matrix navigation. IEEE Trans. Vis. Comp. Graph. **14**(6), 1148–1539 (2008)
9. Fanea, E., Carpendale, M.S.T., Isenberg, T.: An interactive 3D integration of parallel coordinates and star glyphs. In: InfoVis, pp. 149–156 (2005)
10. Friendly, M.: Corrgrams: exploratory displays for correlation matrices. Am. Stat. **56**(4), 316–324 (2002)
11. Geng, Z., Peng, Z., Laramee, R.S., Roberts, J.C., Walker, R.: Angular histograms: frequency-based visualizations for large, high dimensional data. IEEE Trans. Vis. Comput. Graph. **17**(12), 2572–2580 (2011)

12. Harrison, L., Yang, F., Franconeri, S., Chang, R.: Ranking visualizations of correlation using Weber's law. IEEE Trans. Vis. Comput. Graph. **20**(12), 1943–1952 (2014)
13. Hartigan, J.A.: Printer graphics for clustering. JSCS **4**(3), 187–213 (1975)
14. Heinrich, J., Stasko, J., Weiskopf, D.: The parallel coordinates matrix. In: EuroVis - Short Papers, pp. 37–41 (2012)
15. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. In: Eurographics STAR, pp. 95–116 (2013)
16. Holten, D., van Wijk, J.J.: Evaluation of cluster identification performance for different PCP variants. In: EuroVis, vol, 29, no 3 (2010)
17. Hong, X., Wang, C.X., Thompson, J.S., Allen, B., Malik, W.Q., Ge, X.: On space-frequency correlation of UWB MIMO channels. IEEE Trans. Veh. Tech. **59**(9), 4201–4213 (2010)
18. Huang, T.H., Huang, M.L., Zhang, K.: An interactive scatter plot metrics visualization for decision trend analysis. In: Conference on Machine Learning, Applications, pp. 258–264 (2012)
19. Inselberg, A.: The plane with parallel coordinates. Vis. Comput. **1**(2), 69–91 (1985)
20. Jarrell, S.B.: Basic Statistics. W.C. Brow Comm, Reading (1994)
21. Johansson, J., Ljung, P., Jern, M., Cooper, M.: Revealing structure in visualizations of dense 2D and 3D parallel coordinates. Inf. Vis. **5**(2), 125–136 (2006)
22. Kay, M., Heer, J.: Beyond Weber's law a second look at ranking visualizations of correlation. IEEE Trans. Vis. Comput. Graph. **22**(1), 469–478 (2016)
23. Li, J., Martens, J.B., van Wijk, J.J.: Judging correlation from scatterplots and parallel coordinate plots. InfoVis **9**, 13–30 (2010)
24. Mackinlay, J.: Automating the design of graphical presentations of relational information. ACM Trans. Graph. **5**(2), 110–141 (1986)
25. Magnello, E., Vanloon, B.: Introducing Statistics: A Graphic Guide. Icon Books, London (2009)
26. Sharma, R.K., Wallace, J.W.: Correlation-based sensing for cognitive radio networks bounds and experimental assessment. IEEE Sens. J. **11**(3), 657–666 (2011)
27. Tominski, C., Abello, J., Schumann, H.: Axes-based visualizations with radial layouts. In: ACM symposium on Applied computing, pp. 1242–1247. ACM (2004)
28. Wang, J., Zheng, N.: A novel fractal image compression scheme with block classification and sorting based on pearsons correlation coefficient. IEEE Trans. Image Process. **22**(9), 3690–3702 (2013)
29. Wattenberg, M.: Visual exploration of multivariate graphs. In: SIGCHI, CHI 2006, pp. 811–819 (2006)
30. Wilkinson, L., Anand, A., Grossman, R.L.: Graph-theoretic scagnostics. InfoVis. **5**, 21 (2005)
31. Xu, W., Chang, C., Hung, Y.S., Fung, P.C.W.: Asymptotic properties of order statistics correlation coefficient in the normal cases. IEEE Trans. Signal Process. **56**(6), 2239–2248 (2008)
32. Yu, S., Zhou, W., Jia, W., Guo, S., Xiang, Y., Tang, F.: Discriminating DDoS attacks from flash crowds using flow correlation coefficient. IEEE Trans. PDS **23**(6), 1073–1080 (2012)
33. Zhou, H., Cui, W., Qu, H., Wu, Y., Yuan, X., Zhuo, W.: Splatting lines in parallel coordinates. Comput. Graph. Forum **28**(3), 759–766 (2009)
34. Zhou, H., Yuan, X., Qu, H., Cui, W., Chen, B.: Visual clustering in parallel coordinates. In: Computer Graphics Forum (2008)