InterAxis: Steering Scatterplot Axes via Observation-Level Interaction

IEEE VAST 2015

Hannah Kim¹, Jaegul Choo², Haesun Park¹, Alex Endert¹
Georgia Tech¹, Korea University²

October 28th 2015
How would you visualize high-dimensional data?
## Example - Car Data

<table>
<thead>
<tr>
<th>Vehicle Name</th>
<th>Retail Price</th>
<th>Dealer Cost</th>
<th>Engine Size (l)</th>
<th>Cyl</th>
<th>HP</th>
<th>City MPG</th>
<th>Hwy MPG</th>
<th>Weight</th>
<th>Wheel Base</th>
<th>Len</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acura 3.5 RL 4dr</td>
<td>43,755</td>
<td>39,014</td>
<td>3.5</td>
<td>6</td>
<td>225</td>
<td>18</td>
<td>24</td>
<td>3,880</td>
<td>115</td>
<td>197</td>
<td>72</td>
</tr>
<tr>
<td>Acura 3.5 RL w/Navigation 4dr</td>
<td>46,100</td>
<td>41,100</td>
<td>3.5</td>
<td>6</td>
<td>225</td>
<td>18</td>
<td>24</td>
<td>3,893</td>
<td>115</td>
<td>197</td>
<td>72</td>
</tr>
<tr>
<td>Acura MDX</td>
<td>36,945</td>
<td>33,337</td>
<td>3.5</td>
<td>6</td>
<td>265</td>
<td>17</td>
<td>23</td>
<td>4,451</td>
<td>106</td>
<td>189</td>
<td>77</td>
</tr>
<tr>
<td>Porsche 911 GT2 2dr</td>
<td>192,465</td>
<td>173,560</td>
<td>3.6</td>
<td>6</td>
<td>477</td>
<td>17</td>
<td>24</td>
<td>3,131</td>
<td>93</td>
<td>175</td>
<td>72</td>
</tr>
<tr>
<td>Acura RSX Type S 2dr</td>
<td>23,820</td>
<td>21,761</td>
<td>2</td>
<td>4</td>
<td>200</td>
<td>24</td>
<td>31</td>
<td>2,778</td>
<td>101</td>
<td>172</td>
<td>68</td>
</tr>
<tr>
<td>Acura TL 4dr</td>
<td>33,195</td>
<td>30,299</td>
<td>3.2</td>
<td>6</td>
<td>270</td>
<td>20</td>
<td>28</td>
<td>3,575</td>
<td>108</td>
<td>186</td>
<td>72</td>
</tr>
<tr>
<td>Acura TSX 4dr</td>
<td>26,990</td>
<td>24,647</td>
<td>2.4</td>
<td>4</td>
<td>200</td>
<td>22</td>
<td>29</td>
<td>3,230</td>
<td>105</td>
<td>183</td>
<td>69</td>
</tr>
<tr>
<td>Audi A4 1.8T 4dr</td>
<td>25,940</td>
<td>23,508</td>
<td>1.8</td>
<td>4</td>
<td>170</td>
<td>22</td>
<td>31</td>
<td>3,252</td>
<td>104</td>
<td>179</td>
<td>70</td>
</tr>
<tr>
<td>Audi A4 3.0 4dr</td>
<td>31,840</td>
<td>28,846</td>
<td>3</td>
<td>6</td>
<td>220</td>
<td>20</td>
<td>28</td>
<td>3,462</td>
<td>104</td>
<td>179</td>
<td>70</td>
</tr>
</tbody>
</table>

Scatterplot / Scatterplot Matrix

Dimension Reduction: Linear

- Principal Component Analysis (PCA)
  - Linear combination of attributes that maximizes the variance

- Linear Discriminant Analysis (LDA)
  - Linear combination of attributes that separates classes well

Source: http://sebastianraschka.com/Articles/2014_python_lda.html
Dimension Reduction: Nonlinear

- Multidimensional Scaling (MDS)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Force-Directed Layout

Interpretability Issue - Axes do not have clear meaning or are not defined at all.

Another Issue - Interactivity

- Dimension reduction techniques are generally automated.
- Interaction with dimension reduction techniques is not easy.
- To make adjustments, one has to try different parameters and check visualization results.

Hard to reflect user intentions!
Can scatterplot axes be interpretable and interactive?
With InterAxis...

- Users pick two (or more) data items with semantic meanings
  - Ones they like vs. ones they don’t like
  - Ones they are interested in vs. ones they don’t care
  - and so on …

- Then, InterAxis automatically calculates an axis that reflects the semantic meanings.
  - Data items similar to the first group has high values and data items similar to the second group has low values.
  - Each feature’s contribution(weight) to the axis is visualized in a bar chart.
Observation-Level Interaction

- Direct interactions with visual objects to reflect user intent
  - **Data-level:** data objects (dots)
    Find data items that quantify subjective preferences
  - **Feature-level:** features (bars)
    Directly manipulate contributions/weights of features that represent an axis
Live Demo - Car Data

http://va.gatech.edu/projects/interaxis/
Behind The Scenes (1)

- Select two semantic groups of data items
- Find centroids of each group
- Subtract one from the other to get a projection vector
- Normalize the vector

\[ T_x = \frac{1}{n_{x,h}} \sum_{i=1}^{n_{x,h}} a_{i}^{x,h} - \frac{1}{n_{x,l}} \sum_{i=1}^{n_{x,l}} a_{i}^{x,l}. \]
Behind The Scenes (2)

- Project origin (0,0) onto the axis vector
- Project data items onto the axis vector
- New coordinate value is the distance from the projected origin to the projected data point.

\[ (T_x^T a_i, T_y^T a_i) \]
### Discussion: Beyond Linear Models (1)

<table>
<thead>
<tr>
<th>Linear</th>
<th>Automated</th>
<th>User-Driven</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCA</td>
<td>InterAxis</td>
</tr>
<tr>
<td>Non-linear</td>
<td>Manifold learning (e.g., LLE, Isomap)</td>
<td>?</td>
</tr>
</tbody>
</table>

- **Automated**: PCA
- **User-Driven**: InterAxis
- **Non-linear** Manifold learning (e.g., LLE, Isomap)
  - ?
InterAxis represents an axis as a weighted linear combination of data attributes.

However, semantic meanings are not necessarily “linear”.

E.g., drawing a curve that means a progression of frowning to smiling.

Source: Saul and Roweis, locally linear embedding (LLE) paper (JMLR’13)
Discussion: Handling Sparse Data

- Few non-zero entries for an attribute (or for an item)
- Common in significantly high-dimensional data
  - E.g. text, image, gene expression data
- We have to assign more data items to specify an axis.
- One solution is to aggregate multiple attributes into a group.
Summary

- We introduce InterAxis, a visual analytics technique that enables users to
  - Directly define and manipulate axes via observation-level interactions
  - Understand data attributes that quantify subjective preferences