Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, University of Vienna
Tamara Munzner, University of British Columbia
Melanie Tory, University of Victoria
High-dimensional Data

<table>
<thead>
<tr>
<th>length</th>
<th>weight</th>
<th>speed</th>
<th>hp</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

highdim
Dimension Reduction (DR)

dimension reduction

e.g., using PCA

<table>
<thead>
<tr>
<th></th>
<th>length</th>
<th>weight</th>
<th>speed</th>
<th>hp</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DR

<table>
<thead>
<tr>
<th></th>
<th>sporty</th>
<th>handling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Visualizing DR Data

<table>
<thead>
<tr>
<th>sporty</th>
<th>handling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car 1</td>
<td></td>
</tr>
<tr>
<td>Car 2</td>
<td></td>
</tr>
<tr>
<td>Car 3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

lowdim

Visualization

2D Scatterplot

interactive 3D Scatterplot

Scatterplot Matrix (SPLOM)
Which visual encoding technique to use for visualizing DR data?

2D, 3D, SPLOM?
Related Work

General abstract data

• 3D often inappropriate

Chalmers: Using a landscape metaphor to represent a corpus of documents [COSIT’93]
Cockburn and McKenzie: An evaluation of cone trees [British Conf. on HCI’00]
Cockburn and McKenzie: Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments [CHI’02]
Newby: Empirical study of a 3D visualization for information retrieval tasks [Intelligent Information Systems’02]
Tory et al.: Spatialization design: comparing points and landscapes [InfoVis’07]
Tory et al.: Comparing dot and landscape spatializations for visual memory differences [InfoVis’09]
Westerman and Cribbin: Mapping semantic information in virtual space: dimensions, variance and individual differences [IJHCS’00]

DR data

• 3D **is** used in certain domains
• No studies on scatterplot choices for DR data
Contributions

1. Data Study
   • in-depth analysis of 816 scatterplots
   • task: visual cluster verification
Contributions

1. Data Study
   • qualitative analysis of 816 scatterplots
   • task: visual cluster verification

2. Workflow Model

(see paper)
2 part project
**Same method/base data:**
data study with same 816 scatterplots
**Same method/base data:**
data study with same 816 scatterplots

**Different data gathering/analysis:**
- qualitative coding
- quantitative data

**Different goals/contributions:**
- taxonomy of visual cluster separation factors
- evaluation of automatic class separation measures
- Comparing visual encoding choices: 2D, 3D, and SPLOM
Method
Data Study

Many Scatterplots

2 human expert coders
Data Study

Reasons:

• data characteristics outweigh user differences
• need for reliable cluster separation judgement

Sedlmair et al.: A taxonomy of visual cluster separation factors [EuroVis’12]
75 pre-classified datasets

- real (31)
- Gaussian (16)
- entangled (24)
- grid (4)
- synthetic
75 pre-classified datasets

4 DR techniques

• PCA (linear)
• Robust PCA (linear)
• Glimmer MDS (non-linear)
• t-SNE (non-linear)
75 pre-classified datasets
4 DR techniques
3 visual encodings

→ 816 Plots

SPLOM:
3 - 7 dim.
2 human expert coders

- inspect all 816 Plots
- judge all clusters:

  5 = nicely separated
  4 ...
  3 ...
  2 ...
  1 = not separated
2 human expert coders

- inspect all 816 Plots
- judge all clusters:

5 = nicely separated
4 ...
3 ...
2 ...
1 = not separated

Class judgments / coder
~80 hours coding / coder
Judging Reliability

• high inter-coder reliability (Krippendorff’s alpha = 0.86)
• echoing previous findings

Lewis et al.: Human cluster evaluation and formal quality measures: a comparative study [CogSci’12]
Data Analysis & Results
Cost Assumption

2D < SPLOM < 3D

• Based on rich body of previous work*

* previous work:

Drawbacks of 3D

Chalmers: Using a landscape metaphor to represent a corpus of documents [COSIT'93]

Cockburn and McKenzie: An evaluation of cone trees [British Conf. on HCI'00]

Cockburn and McKenzie: Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments [CHI'02]


Tory et al.: Spatialization design: comparing points and landscapes [InfoVis'07]

Tory et al.: Comparing dot and landscape spatializations for visual memory differences [InfoVis'09]

Westerman and Cribbin: Mapping semantic information in virtual space: dimensions, variance and individual differences [IJHCS'00]

Interaction Costs

Lam: A framework of interaction costs in information visualization [InfoVis'08]

Van Wijk: Views on visualization [TVCG'06]
Cost Assumption

2D < SPLOM < 3D

• Based on rich body of previous work

Reasons:

• 2D (low): static, directly visible
• SPLOM (medium): switching attention between views
• 3D (high): interaction to resolve occlusions
Cost Assumption

• Use a higher cost visual encoding **only** if it provides notably better class separation

• Use 2D if “good enough”, if not then SPLOM, then 3D
Data Analysis

1. Heatmaps Approach
   • reveals a lot of the details

2. Statistical Analysis
   • confirms heatmap analysis
   • see paper
Base heatmaps

Showing averaged scores of two coders
4 DR techniques

- PCA
- robust PCA
- glimmer MDS
- t-SNE

row = scatterplot
Delta Heatmaps: Cell-wise difference

Delta Heatmap

A  better

B better
Within-DR
SPLOM vs. 2D

which is better?

SPLOM_{PCA} vs. 2D_{PCA}
SPLOM vs. 2D

- PCA
- robust PCA
- glimmer MDS
- t-SNE

SPLOM colors:
- substantially
- noticeable
- marginal
- same
- marginal
- noticeable
- substantially

2D colors:

Michael Sedlmair, University of Vienna
SPLOM vs. 2D

PCA
robust PCA
glimmer MDS
t-SNE

SPLOM_{robPCA}
2D_{robPCA}

real
Gaussian
entangled
grid
3D vs.
best of (2D, SPLOM)

which is better?

$3D_{PCA}$

$2D_{PCA}$ or $SPLOM_{PCA}$
3D vs. (2D, SPLOM)

3D

- substantially
- noticeable
- marginal
- same
- marginal
- noticeable
- substantially

2D or SPLOM

Michael Sedlmair, University of Vienna
3D vs. (2D, SPLOM)

- PCA
- robust PCA
- glimmer MDS
- t-SNE

real

Gaussian

entangled

grid

3D_{PCA}  vs.  2D_{PCA}  SPLOM_{PCA}

Michael Sedlmair, University of Vienna
3D vs. (2D, SPLOM)

**PCA**

**robust PCA**

**glimmer MDS**

**t-SNE**

Gaussian

real

entangled

grid

3D

2D

SPLOM

Michael Sedlmair, University of Vienna

data: gauss-n500-10d-5smallIC, synthetic-Gaussian, glimmer MDS
Between-DR
2D vs. best of (2D from other DRs)

which is better?

2D_{PCA}

2D_{robust PCA}

2D_{glimmer MDS}

2D_{t-SNE}
Cross-column differences in 2D base heatmap
Cross-column differences in 2D base heatmap
Cross-column differences in 2D base heatmap
Cross-column differences in 2D base heatmap
2D vs. (2D from other DRs)

no one and only DR

“own” DR’s 2D

“another” DR’s 2D

substantially
noticeable
marginal
same
marginal
noticeable
substantially

Michael Sedlmair, University of Vienna

InfoVis’13
2D vs. (2D from other DRs)

- PCA
- Robust PCA
- GLIMMER
- t-SNE

**t-SNE** good for highly synthetic datasets:

**entangled** (intended to benefit 3D)

**grid**
2D vs. (2D\text{from other DRs})

**same dataset, different DR**

PCA

robust PCA

**glimmer MDS**

t-SNE

real

Gaussian

entangled

grid

2D_{MDS}

vs.

2D_{t-SNE}
SPLOM vs. best of \((2D_{\text{from all DRs}})\)

which is better?

SPLOM_{PCA}
SPLOM vs. (2D\textsubscript{from all DRs})

- PCA
- robust PCA
- glimmer MDS
- t-SNE

2D\textsubscript{PCA}  2D\textsubscript{robPCA}  2D\textsubscript{MDS}  2D\textsubscript{t-SNE}

SPLOM\textsubscript{PCA}

data: industryIndices, real, PCA
3D vs. best of (2D_{from all DRs}, SPLOM)

which is better?

3D_{PCA}
3D vs. (SPLOM\textsubscript{own}, 2D\textsubscript{from all DRs})

- PCA
- robust PCA
- glimmer MDS
- t-SNE

- real
- Gaussian
- entangled

no noticeably better class in 3D

3D
- substantially
- noticeable
- marginal
- same

SPLOM or one of DR’s 2D

from all DRs
Summary
Summary

Which visual encoding to use for dimensionally reduced data?
- 2D, interactive 3D, SPLOM?

Data study
- Heatmap analysis
- Examples
Results

- 3D “better”
- SPLOM “better”
- 2D good enough

“better” = at least one class is notably more separable in SPLOM or 3D
Implications

• **Use 2D:** 2D often good enough
• **Change DR:** if not, change DR technique
• **Then SPLOM:** SPLOM occasionally helps
• **No 3D:** 3D rarely helps and often hurts
Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices

Michael Sedlmair, Tamara Munzner, Melanie Tory
contact: michael.sedlmair@univie.ac.at
project page: http://www.cs.ubc.ca/labs/imager/tr/2013/ScatterplotEval/