INDEPENDENT STUDY REPORT S'25

Bipasha Garg
2022111006

High-Dimensional Data Visualization Techniques

1. Introduction

This report includes the key findings and concepts from reviews of several prominent papers and a thesis focused on visualizing high-dimensional data. The techniques include SpaceMAP, Circle Segments, Heidi Matrix (for R-Tree visualization), and insights from a survey on Parallel Coordinates evaluations.

2. Visualizing High-Dimensional Data

Key challenges identified include:

- **The Crowding Problem:** Placing points in a lower-dimensional space leads to loss of structure and information due to overlap between these points (illustrated by SpaceMAP). Distributions of pairwise distance are affected drastically, concentrating at larger distances in high dimensions, making naive mapping impossible.
- **Loss of Structure:** During reduction or visualization, local neighbourhood information as well as global structure such as clusters or manifolds may be fused, distorted, or completely erased.
- **Curse of Dimensionality:** Distance metrics become differentiable, while the volume of space grows proportionality, exponentially lessening the intuitiveness of "nearness" and increasing complexity of computational analysis.
- **Scalability:** An increased number of data points (N) and dimensions (D) creates challenges in computation, memory, as well as visual clutter.

3. Core Techniques

- **SpaceMAP:** Addresses crowding by explicitly aiming to match the "capacity" (volume) of high- and low-dimensional spaces. It uses the concept of Equivalent Extended Distance (EED) to transform high-dimensional distances *before* calculating similarities. Key features include hierarchical modeling (near/middle/far fields), estimation of intrinsic dimensionality, and a connection to contrastive learning loss functions (similar to UMAP).

- **Circle Segments:** A pixel-per-value technique designed for large, high-dimensional datasets. It maps dimensions to circular segments and arranges data values within them, leveraging color mapping for intuition. Strengths include handling more dimensions than screen-width-limited linear techniques and offering interactive dimension rearrangement.
- **Heidi Matrix (for R-Trees):** Visualizes high-dimensional index structures (like R-Trees) using a 2D matrix layout. Rows/columns are ordered based on the R-Tree hierarchy. Cell color encodes the "closeness" (based on kNN) of point pairs across various subspaces, revealing index structure and subspace relationships simultaneously.
- **Parallel Coordinates (PC) & Extensions:** The PC Evaluation survey provides context on a standard technique. PC represents dimensions as parallel vertical axes and data points as lines connecting values across axes. The survey highlights the importance of user-centered evaluation for variations (axis layouts, clutter reduction) and practical applications.

4. Key Concepts across Techniques

- **Addressing Crowding:** SpaceMAP deals with this directly via EED and space capacity matching. Other techniques address it implicitly, for example, through layout (Circle Segments circumvents the linear screen limit constraints) or focus (Heidi Matrix utilizes points of structural relations rather than merely proximity based relations). The PC survey mentions clutter as one of the key areas of evaluation.
- **Handling Hierarchy and Structure:** SpaceMAP the hierarchy of data structures (sub-manifolds) explicitly. Heidi Matrix is specially made to depict hierarchical structures of an R-Tree index. This idea is important for the BEADS data which is described as having potential hierarchical structures (clusters/sub-manifolds).
- **Subspace Awareness:** Heidi Matrix portrays the best subspace awareness because of the change of relationship of the points in different subspaces (cell color coding) in it. MyViz, representing the subspaces in rings and quadrants in sectors, aims at depicting information particular to different subspaces.
- **Intrinsic vs. Observed Dimensionality:** SpaceMAP incorporates estimating the intrinsic dimensionality (the “true” complexity that one assumes is carefully observed to be lower than the observable one) which is the case in manifold learning.
- **Similarity Metrics and Distance Transformations:** Transforming distances to similarly be measured differently (difference via EED) is what SpaceMAP does in an original way. Both SpaceMAP and UMAP (mentioned as similar) use Gaussian-like/exponential similarity functions and contrastive loss functions. Heidi Matrix uses kNN-based closeness in subspaces.
- **Scalability Concerns:**
  - *Computational:* Heidi Matrix has $O(dn^2)$ complexity, problematic for large N. SpaceMAP's complexity wasn't detailed but likely significant.

    - *Memory:* Heidi Matrix requires storing an N x N grid, demanding large memory for large N.
    - *Visual:* High dimensions can lead to visual clutter in Heidi Matrix (many subspaces/colors) and potentially in Circle Segments/MyViz (many segments/sectors, potentially negligible areas for outer rings/higher dimensions). PC also suffers from visual clutter.
- **Interaction and User Control:** Circle Segments emphasizes interactive dimension rearrangement for comparison. The PC Evaluation survey highlights interaction (adding/removing axes) as a positive factor. It is identified as a key area for My Viz development (proportional views, zooming, collapsing rings, polar vs. radial).
- **Intuitiveness and Learnability:** Circle Segments is noted as potentially non-intuitive initially. The PC survey notes that standard PCs are relatively easy to learn. Heidi Matrix's color encoding requires a legend and expertise. My Viz has a similar challenge – the meaning of rings, sectors, colors, and edges needs to be learned.
- **User-Centered Evaluation:** The PC Evaluation paper strongly advocates for empirical, user-centered studies to determine the actual effectiveness and usability of visualization techniques, rather than relying solely on theoretical arguments or algorithm performance.

5. Relevance and Implications for "MyViz"

- **Positioning MyViz:** My Viz is a mixture of different approaches. Like Circle Segments, it uses a circular layout to overcome linear limitations and potentially handle many dimensions. Like Heidi Matrix (conceptually), it uses spatial partitioning (rings for subspaces, sectors for quadrants) to convey structural information. Its goal of showing spatial arrangement relates to the general aims of DR/visualization techniques like SpaceMAP/UMAP/t-SNE.
- **Addressing a few Questions:**
    - ***Clutter**: The PC Review question proposes investigating transparency for edges. My Viz requires strategies to cope with possible clutter from numerous points, dimensions (sectors), and labels.*
    - ***Adjacent Relationships:** The PC review question ("Does My Viz enable identification of relationships between adjacent axes?") requires attention. In MyViz's circular layout, "adjacency" could imply adjacent sectors in a ring or adjacent rings.*
    - ***Intuitiveness:** The original non-intuitiveness of Circle Segments implies MyViz will need to have clear onboarding or tutorials. The possible "cognitive load" (rings, sectors, edge colors, labels) mentioned in the Heidi Matrix review is extremely applicable. A "straight band representation" could make the initial view easier, as implied.*

    - ***Interaction:*** *Enforcing planned interactions (proportional views, zoom, collapse/expand rings, polar view) is consistent with research that flexibility and interaction enhance usability (PC Evaluation, Circle Segments).*
    - ***Scalability:*** *The computational and memory expenses emphasized for Heidi Matrix are cautionary for My Viz. Large N and D performance testing is essential. Visual scalability (ability to handle lots of dimensions/sectors without being unreadable) must also be tested.*
    - ***Dimensionality/Subspace Ordering:*** *The selection of variance/gini index for ordering rings (subspaces) in MyViz is a key design choice, similar to the effect of 'k' and distance metric in Heidi Matrix's kNN.*
    - ***Color Usage:*** *The Circle Segments comment on mapping values to color for intuition is directly relevant. My Viz's use of edge color for labels must be designed carefully to stay clear.*
    - ***Evaluation:*** *The PC survey's strong focus on user-centered evaluation highlights the importance of quantitative and qualitative user tests for My Viz to evaluate readability, interpretability, and task performance.*

7. Synthesized Challenges and Considerations

- **Trade-off: Intuitiveness vs. Information Density:** More advanced approaches (SpaceMAP, Heidi Matrix, possibly MyViz) capture more information, but as with most advanced techniques, they often require more user effort for interpretation and training compared to more rudimentary methods.
- **Scalability Remains Key:** Dealing with large N and D presents continuous challenges such as computational, memory, visual, and even spatial considerations. No single technique reviewed seems to address all the scalability issues perfectly.
- **The Importance of Interaction:** Static visualizations are impractical and limited. Beside basic exploration, interactivity such as filtering and dynamic parameter changing like ‘k’ in kNN or dimension reordering, is vital for practical implementation.
- **No One-Size-Fits-All:** The answer to this question is largely contingent upon the characteristics of the data (dimensionality, structure, size) and the specific analysis task (exploratory data analysis, cluster identification, outlier detection, index diagnostics).
- **Rigorous Evaluation is Non-Negotiable:** Claims of effectiveness, as pointed out in the PC survey, requires studies which measure performance using relevant tasks to be performed by real users.

8. Conclusion

This comparative analysis of SpaceMAP, Circle Segments, Heidi Matrix, and PC assessments gives a rich context of methods, ideas, and issues in high-dimensional visualization. Main points are the direct management of space capacity by SpaceMAP, Circle Segments' pixel-based scalability, Heidi Matrix's subspace-aware structural visualization, and the general necessity of user-centered evaluation and interaction. These observations directly guide MyViz's design, development, and validation approach, specifically scalability, intuitiveness, interaction design, and applicability to hierarchically organized data. Overcoming the challenges noted, in particular scalability and cognitive load, and capitalizing on interactive subspace exploration strengths are important.

9. References

- Zu, Xinrui, and Qian Tao. *SpaceMAP: Visualizing High-dimensional Data by Space Expansion*.
- Ankerst, Mihael, Daniel A. Keim, and Hans-Peter Kriegel. *'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets*.
- *Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research*.
- Shradha's Thesis