Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

*Volume 33 (2014), Number 3*

# Supplementary Material: Distortion-Guided Structure-Driven Interactive Exploration of High-Dimensional Data

S. Liu[1], B. Wang[1], P.-T. Bremer[2] and V. Pascucci[1]

[1]Scientific Computing and Imaging Institute, University of Utah
[2]Lawrence Livermore National Laboratory

## 1. Data Standardization

High-dimensional data may be pre-processed with a standardization process. Since different parameters may be measured on different scales and the range of values may differ from each dimension, some parameters may dominate the results of the analysis. Various methods exist for data standardization [Gow85, MC88, KR90]. Several methods under considerations are:

- Z-score scaling: values $V$ of each dimension are recomputed as $V - mean(V)/std(V)$; therefore all input parameters have the same mean (0) and standard deviation (1) but different ranges.
- $[0,1]$-Scaling: $V$ is recomputed as $V - \min(V)/(\max V - \min V)$. The input variables have the same ranges but different means and standard deviations.

## 2. Performance Data

To evaluate the performance of our tool during interactions, we provide a detailed report regarding the performance of online computations of point-wise distortion measures. The performance data is generated on a desktop machine equipped with an Intel Core i5 2.6GHz CPU and 8GB of memory. In Table 1, we provide the number of dimensions and point count for each dataset used in the performance analysis.

| Dataset | # of points | # of dimensions |
|---|---|---|
| Parabola | 900 | 3 |
| Combustion | 2796 | 11 |
| Nuclear | 608 | 106 |

**Table 1:** *Testing datasets.*

During the interactive data exploration, DR results are not recomputed. The only computation required is the recalculation of point-wise distortion measures after data manipulation. In Table 2, we list the performance numbers (unit: second) for DR-independent point-wise distortion measures. RDD stands for robust distance distortion and KDE stands for kernel density estimation distortion. Since those distortion measures are DR-independent, we use PCA as the DR method.

| Dataset | Stress | RDD | KDE | Co-ranking |
|---|---|---|---|---|
| Parabola | 0.0171 | 0.0225 | 0.0780 | 0.0504 |
| Combustion | 0.1538 | 0.0946 | 0.6656 | 0.3705 |
| Nuclear | 0.0115 | 0.0097 | 0.0640 | 0.0305 |

**Table 2:** *Timing for DR-independent distortion measures under PCA.*

| Dataset | PCA | MDS | LE | LLE |
|---|---|---|---|---|
| Parabola | $2.33e^{-4}$ | $1.10e^{-2}$ | $8.27e^{-4}$ | $6.54e^{-4}$ |
| Combustion | $1.33e^{-3}$ | $7.40e^{-2}$ | N/A | $4.14e^{-2}$ |
| Nuclear | $7.13e^{-4}$ | $3.18e^{-4}$ | $2.14e^{-2}$ | $1.45e^{-3}$ |

**Table 3:** *Timing for DR-dependent distortion measures.*

Table 3 lists performance numbers (unit: second) for DR-dependent point-wise distortion measures, for typical DR techniques: PCA, MDS, LE and LLE. The missing data (N/A) is the result of the failure of Laplacian Eigenmap (LE) algorithm to build a connected graph with a reasonably large neighborhood size.

From the above tables, we observe that the slowest operation takes less than 0.7 second to complete, and most operations use only a fraction of a second. For the datasets we have tested, realtime interactivity is guaranteed in our framework for datasets of moderate sizes. System scalability for large-scale datasets remains a challenge.

## References

[Gow85] GOWER J. C.: Measures of similarity, dissimilarity, and distance. In *Encyclopedia of Statistical Sciences*, Kotz S., Johnson N., Read C., (Eds.), vol. 5. John Wiley & Sons, 1985, pp. 397–405. 1

[KR90] KAUFMAN L., ROUSSEEUW P. J.: *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons, 1990. 1

[MC88] MILLIGAN G. W., COOPER M. C.: A study of standardization of variables in cluster analysis. *J. Classif. 5* (1988), 181–204. 1