FINDING VISUALIZATIONS RELEVANT TO THE USER

Seppo Virtanen[†] [‡], Homayun Afrabandpey [‡] and Samuel Kaski [‡]

[†]Department of Statistics, University of Warwick

[‡]Helsinki Institute for Information Technology HIIT , Department of Computer Science , Aalto University

INTRODUCTION

- We introduced a method for learning visualizations relevant to the user by multi-view latent variable factorization
- Assuming user interaction data, the goal of multi-view visualization is to identify which aspects in the primary data support the user's input and which aspects of the user's potentially noisy input have support in the primary data
- The proposed method exploits two sources (views) of information (primary data and the user interactions) to identify which aspects in the two views are related and which are specific to only one of them.

APPROACH

$D = [d]_{ij}$ and $F = [f]_{ij}$ are two relational count data sets representing similarities between pairs of N samples, $\{x_i\}_{i=1}^N$, from two different views (D for data view and F for user view). The two views, D and F, are modeled with distributions p and q, respectively.

$$p(D, F, \Theta) \propto \prod_{i=1,j>i}^{N} p_{i,j}^{\tilde{d}_{i,j}} \prod_{i',j' \in \mathcal{O}} q_{i',j'}^{\tilde{f}_{i',j'}}$$

Learning Algorithm:

$$p_{i,j} \propto \exp(-||z_i - z_j||^2 - ||z_i^{(D)} - z_j^{(D)}||^2)$$

$$q_{i,j} \propto \exp(-||z_i - z_j||^2 - ||z_i^{(F)} - z_j^{(F)}||^2)$$

By defining $y_i = [z_i, z_i^{(D)}, z_i^{(F)}]$:

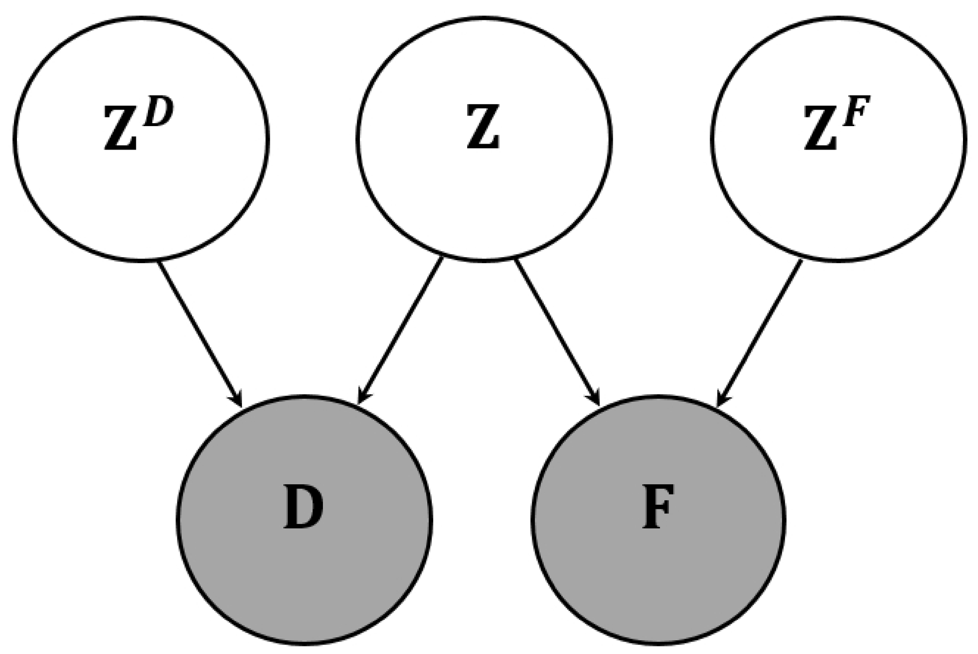
$$p_{i,j} \propto \exp(-(y_i - y_j)^T W^{(D)} W^{(D)^T} (y_i - y_j))$$

$$q_{i,j} \propto \exp(-(y_i - y_j)^T W^{(F)} W^{(F)^T} (y_i - y_j))$$

Locations on the dispaly, y_i, can be estimated by maximizing the following log-likelihood function:

$$\mathcal{L} = \lambda \sum_{i=1,j>i}^{N} \tilde{d}_{i,j} \log p_{i,j} + (1-\lambda) \sum_{i',j' \in \mathcal{O}} \tilde{f}_{i',j'} \log q_{i',j'}$$

GRAPHICAL MODEL



Gray and white nodes depict observed and hidden variables, respectively. The $\mathbf{Z}^D$, $\mathbf{Z}$, and $\mathbf{Z}^F$ are matrices containing all primary-data-specific latent variables ($\mathbf{z}_i^{\mathcal{D}}$), shared latent variables ($\mathbf{z}_i$), and user-data-specific latent variables ($\mathbf{z}_i^{\mathcal{F}}$), respectively. In more detail, the entry d_{ij} of $\mathbf{D}$ depends on the rows i and j (shared vectors $\mathbf{z}_i$ and $\mathbf{z}_j$) of **Z**, and the rows i and j of $\mathbf{Z}^D$; the dependencies for $f_{i,j}$ are analogous.

EXPERIMENTAL RESULTS

Assume a user is interested in grouping of data points. The auxiliary data contains similarity assessments, some of which have support in the primary data and some not.

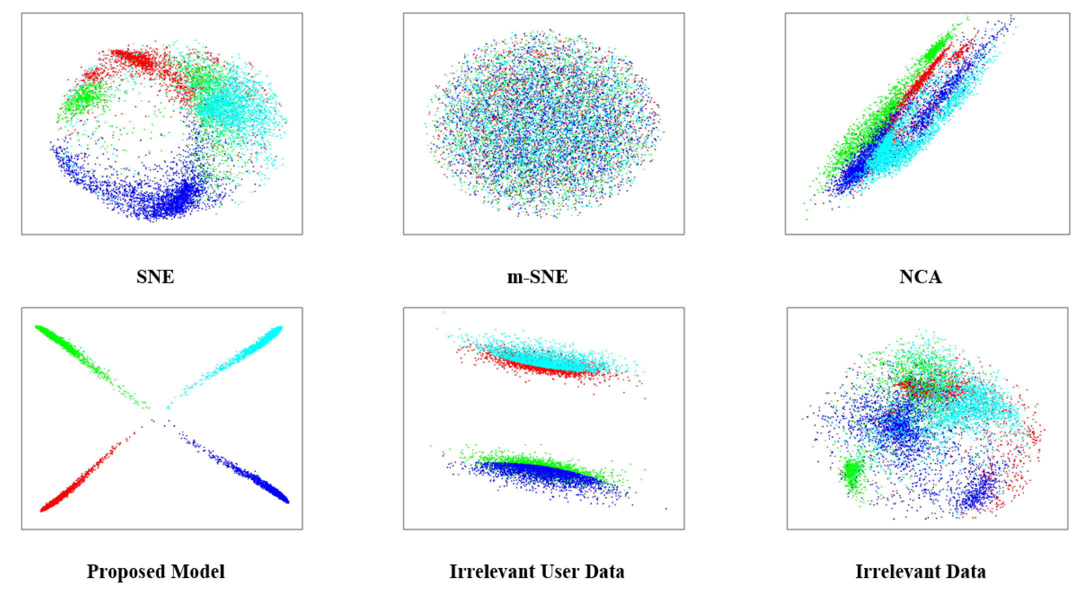


Figure 1: Visualizations obtained by different methods on RCV1 data set

Quantitative Comparison:

Methods		Data Set: Scientific Articles	RCV1	Heart Disease
SNE		60.66%	20.27%	52.15%
m-SNE		62.56%	65.13%	42.21%
NCA		33.18%	19.51%	29.7%
Proposed Method	$K = 6$	9.9%	2.56%	46.20%
	$K = 8$	**0.47%**	**0.76%**	22.11%
	$K = 10$	11.37%	6.85%	**1.65%**

Table 1: Generalization error of k-NN classifier with $k = 5$ on low dimensional representations

Multi-Labelling Case Study:

Assume two users are interested in different aspects of same data and give different feedbacks (labellings). In our experiment, one user is interested in the age of some abalones and the other user is interested in the sex of the abalones.

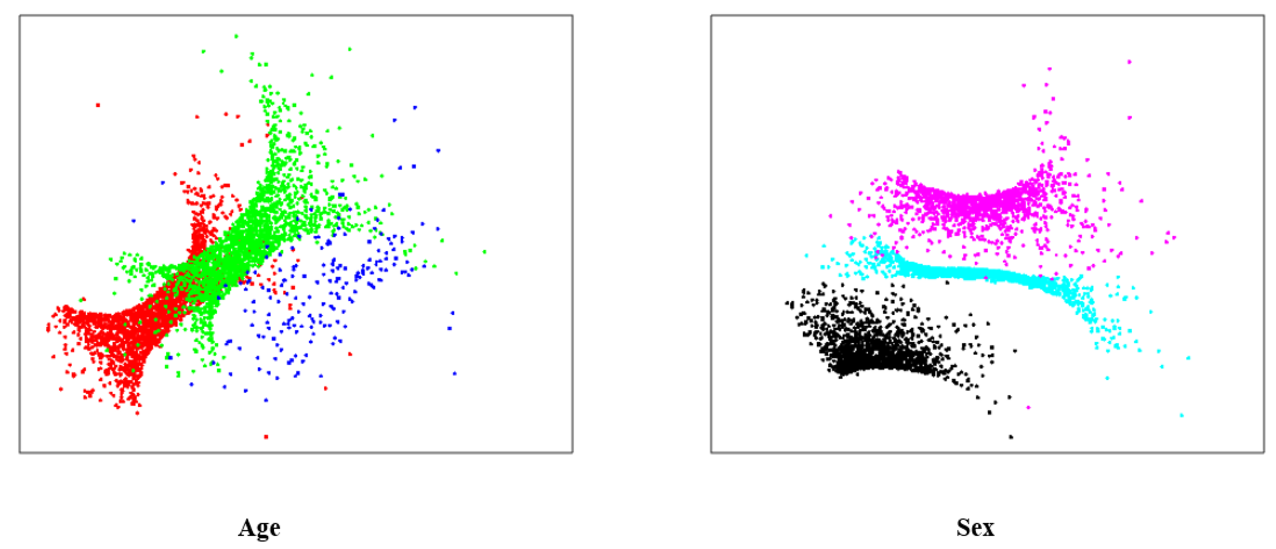


Figure 2: Visualizations of Abalone data for two different users interested in different aspects of the data

CONCLUSION

- We have presented a statistical principle to identify and visualize aspects of data relevant to the user by exploiting statistical relations found between the primary data, and user-provided auxiliary data
- A main future goal is to use similar technique in interactive visualization where user interaction data will be measured all the time and visualization needs to react faster

CONTACT INFORMATION

Web www.research.cs.aalto.fi/pml

Email s.virtanen@warwick.ac.uk
homayun.afrabandpey@aalto.fi
samuel.kaski@aalto.fi

Funding: Academy of Finland (Finnish Center of Excellence in Computational Inference COIN)

REFERENCES

[1] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, pages 833–840, 2002.

[2] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *The Journal of Machine Learning Research*, 14(1):965–1003, 2013.