VISUALIZATIONS RELEVANT TO THE USER BY MULTI-VIEW LATENT VARIABLE FACTORIZATION

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

ABSTRACT

A main goal of data visualization is to find, from among all the available alternatives, mappings to the 2D/3D display which are relevant to the user. Assuming user interaction data, or other auxiliary data about the items or their relationships, the goal is to identify which aspects in the primary data support the user's input and, equally importantly, which aspects of the user's potentially noisy input have support in the primary data. For solving the problem, we introduce a multi-view embedding in which a latent factorization identifies which aspects in the two data views (primary data and user data) are related and which are specific to only one of them. The factorization is a generative model in which the display is parameterized as a part of the factorization and the other factors explain away the aspects not expressible in a two-dimensional display. Functioning of the model is demonstrated on several data sets.

***Index Terms*—** Data visualization, latent factor models, manifold embedding, multi-view learning

1. INTRODUCTION

In the machine learning community there has been a strong trend in developing methods for non-linear dimensionality reduction and manifold embedding, that is, finding lower-dimensional manifolds within high-dimensional data spaces. Examples of these methods include Laplacian eigenmap [1], Isomap [2], locally linear embedding [3] and stochastic neighbor embedding [4]. For a comparison of several methods see [5]. Different methods aim at preserving different geometric properties, but common to all is that they produce a lower-dimensional output where similar data points are located closer together than dissimilar data points.

The low-dimesional projections can be used to construct scatterplots of the data, to fulfill the constantly increasing need for data visualizations across a wide range of applications. If the dimensionality of the data manifold is two, the methods work well for visualization. If it is higher, a simple solution is to choose the output dimensionality to be two, essentially compressing the data manifold. It has turned out that most of the methods are not able to do that well, but by formulating the cost functions in terms of misses and false positives, a desired tradeoff between the two can be optimized for the visualization [6].

Assuming all dimensions of the data manifold are not equally relevant to the user, a better solution is to focus on visualizing the relevant ones. In this paper we assume data are available about the user preferences, from which the visualization can be learned, but the data about preferences may be indirect. A simple example is explicit feedback about groups of data being similar, which can be expressed as cluster memberships or data classes. These types of auxiliary data can contain noise that is structured in the sense of containing classes that are not visible in the primary data. Given the primary data, and auxiliary data about user preferences, the task of finding the relevant aspects of the data is then essentially a two-view learning problem: identify what is statistically shared in the two data views. We additionally will want the shared aspects to be used for the visualization, requiring that the rest of the signals, "structured noise," is explained away. This is what our model is capable of handling.

We build on the popular stochastic neighbor embedding (SNE) method to infer the visualizations. SNE has earlier been formulated for multiple views, as multi-view stochastic neighbor embedding (mSNE) [7]. That work integrated the features into a unified representation but did not yet consider visualization. The model assumed a single set of low-dimensional latent variables which explains all views, and hence cannot directly separate the source-specific "structured noise" from signal. A related model called multiple relational embedding [8] introduced view-specific mappings that can switch latent variables off from the views and hence could implement view-specific "explaining away." They did not aim at separating shared variation from view-specific, however, and did not consider visualization yet either.

In summary, the main contribution of this paper is that we formulate a generative probabilistic model to solve the problem of learning a visualization relevant to the user, operating on two-view data. The first view consists of the primary data, and the second view of auxiliary data collected from the

We thank the Academy of Finland for funding (Finnish Centre of Excellence in Computational Inference Research COIN)

user. The main difference between our model and the existing models is that our model has a set of latent variables, some of which are coordinates of the visualization, and the rest explain away the parts of data not relevant to the user.

2. MODEL

We propose a generative model for two relational count data sets, denoted as $\mathcal{D}$ and $\mathcal{F}$, that represent similarities between pairs of N items. For example, the items can correspond to scientific articles or other documents. We denote the count between items i and j as $d_{i,j}$ for $\mathcal{D}$ and $f_{i,j}$ for $\mathcal{F}$, respectively. We assume the counts are symmetric, that is, $d_{i,j} = d_{j,i}$ for all $i, j \in \{1, \ldots, N\}$. High count indicates strong similarity between two items, whereas low count indicates the two items are less similar. However, we do not assume that counts should be similar between the data sets.

We model the data $\mathcal{D}$ with a distribution p and the user data $\mathcal{F}$ with a distribution q, both defined over pairs of data items. Assuming that user data are available for only a subset $\mathcal{O}$ of data pairs, and the data $\mathcal{D}$ for all pairs, the joint generative distribution of the parameters and data is

$$p(\mathcal{D}, \mathcal{F}, \Theta) \propto \prod_{i=1, j>i}^{N} p_{i,j}^{d_{i,j}} \prod_{i',j' \in \mathcal{O}} q_{i',j'}^{f_{i',j'}}, \qquad (1)$$

where we have collected all parameters to Θ. In the following, for brevity, we limit our notation to the data set $\mathcal{D}$ and variables related to it, noting that similar constructions apply for the $\mathcal{F}$, due to symmetry. We then normalize the observed counts to a distribution over the items denoted by $\tilde{d}$ and $\tilde{f}$, and use the mean-normalized log-likelihood.

2.1. Data visualization and two-view learning

For learning the visualization, we introduce three vector-valued (factorised) latent variables for each item. The variables $\mathbf{z}$ are shared by the two views, and the $\mathbf{z}^{(\mathcal{D})}$ and $\mathbf{z}^{(\mathcal{F})}$ are specific to their respective view. With these latent variables, we construct the latent (mean data) distributions as

$$p_{i,j} \propto \exp(-||\mathbf{z}_i - \mathbf{z}_j||^2 - ||\mathbf{z}_i^{(\mathcal{D})} - \mathbf{z}_j^{(\mathcal{D})}||^2), \qquad (2)$$
$$q_{i,j} \propto \exp(-||\mathbf{z}_i - \mathbf{z}_j||^2 - ||\mathbf{z}_i^{(\mathcal{F})} - \mathbf{z}_j^{(\mathcal{F})}||^2),$$

where $||\cdot||$ denotes the Euclidean distance. The idea is that the shared latent variables capture dependencies (common joint variation) between the two data sets, whereas the data set-specific latent variables capture non-shared variation for each data set. Figure 1 shows a graphical representation of our proposed model where, for simplicity, we have collected the latent variable vectors into matrices to more clearly show the main dependencies.

In this parameterization, distances between the latent variables for any two items are proportional to pair-wise dissimilarity between them. Assuming a large $p_{i,j}$, the model prefers keeping the distance between latent points i and j small. In contrast, a small $p_{i,j}$ does not affect the cost function as strongly. This insight underlies stochastic neighbor embedding and related further developments [4, 6, 9].

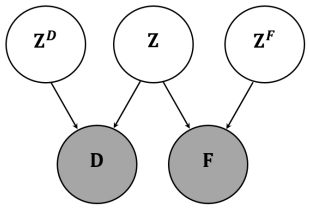


Fig. 1: Graphical model for the two-view latent variable model. Gray and white nodes depict observed and hidden variables, respectively. The $\mathbf{Z}^{\boldsymbol{D}}$, $\mathbf{Z}$, and $\mathbf{Z}^{\boldsymbol{F}}$ are matrices containing all primary-data-specific latent variables ($\mathbf{z}_i^{\mathcal{D}}$), 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 $\mathbf{Z}$, and the rows i and j of $\mathbf{Z}^{\boldsymbol{D}}$; the dependencies for $f_{i,j}$ are analogous.

In this work, we assume the shared latent variables are 2D and use them as visualisation coordinates. Assuming the view-specific latent variables have sufficient modelling flexibility to explain view-specific variation, the shared latent variables will capture the most significant commonalities between the data sets.

2.2. Group-sparse formulation

The model can be written more compactly by concatenating the latent variables together,

$$\mathbf{y}_i = [\mathbf{z}_i, \mathbf{z}_i^{(\mathcal{D})}, \mathbf{z}_i^{(\mathcal{F})}],$$

and introducing data set-specific binary indicator variables $b_k^{(\mathcal{D})}$ and $b_k^{(\mathcal{F})}$, for $k = 1, \ldots, K$, that represent the latent variables either as active (one) or non-active (zero) in the corresponding data set. We note that the explicit factorization of the latent space in Equation 2 is useful for understanding the structure of the multi-view model and the role the different latent variables play. However, learning with a model that has fixed factorized latent variables induces severe identifiability problems for any local learning algorithm. An approach to alleviate this problem is to assume a common latent space formulation and relax the binary indicator variables to continuous variables $\mathbf{W}^{(\mathcal{D})}$ and $\mathbf{W}^{(\mathcal{F})}$, respectively. Assuming sparsity for the indicator variables, during learning some of them will approach zero, shutting off the corresponding view.

Thus, we re-parameterize

$$p_{i,j} \propto \exp(-(\mathbf{y}_i - \mathbf{y}_j)^T \mathbf{W}^{(\mathcal{D})} \mathbf{W}^{(\mathcal{D})^T} (\mathbf{y}_i - \mathbf{y}_j)), \qquad (3)$$

where $\mathbf{y}_i$ for $i = 1, \ldots, N$ are the (concatenated) K-dimensional latent variables and $\mathbf{W}^{(\mathcal{D})}$ is a $K \times K$ diagonal

matrix (we make a similar modification for q with $\mathbf{W}^{(\mathcal{F})}$). We set the first two elements of $\mathbf{W}^{(\mathcal{D})}$ (and $\mathbf{W}^{(\mathcal{F})}$) to unity capturing the shared latent variables, whereas the remaining variables on the diagonal are unconstrained. In the experiments, we show empirically on multiple data sets that the construction correctly captures the dominant shared variation between the two views via the shared latent variables.

For estimating the unobserved variables (locations on the display), we used unconstrained gradient-based optimization to find the maximum likelihood estimate.

2.3. Data setup and visualization

So far, we have not specified what the $\mathcal{D}$ and $\mathcal{F}$ correspond to. In this work, the $\mathcal{D}$ denotes the data the user wants to visualize. They may come directly as observations of similarity between item pairs, forming count data $d_{i,j}$, or they may come as feature vectors $\mathbf{x}_i$, from which the similarities are computed as $\tilde{d}_{i,j} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma^2)$. The $\mathcal{F}$ denotes data provided by the user or measured from the user (more details below); they may come directly as count data $f_{i,j}$, or computed from feature vectors $\mathbf{f}_i$ as for the data $\mathbf{x}_i$.

The first option for the types of data the user may provide, is data about pairwise similarities of the data items. The user data can be collected in an interactive data-analysis session, whereby $f_{i,j}$ would be the number of times the items i and j were considered similar, or derived from categorizations or classifications: $f_{i,j}$ then is the number of classes in which both i and j belong. A particularly handy interactive visualization scenario is where the user indicates a set of data items being similar, which reduces to the (multiple) classification setting. The second option for user data types is that a feature vector is measured from the user during interaction, or the user provides an auxiliary feature vector $\mathbf{f}_i$ for some of the i. They can then be converted to $\tilde{f}_{i,j} = \exp(-||\mathbf{f}_i - \mathbf{f}_j||^2/\sigma_f^2)$.

In both options, the user data are regarded as indirect evidence of what is relevant to the user in the primary data. The key assumption is that aspects of the primary data that have a statistical relationship with the user data are more relevant. The rest of the user input is not relevant to the primary data, and the rest of the primary data is not supported by the user input, and hence is likely to be less relevant to the user.

3. EXPERIMENTAL EVALUATION

We compare our approach to the closest available alternatives; none of them have been designed for precisely the proposed task, but they can still be applied for the task. We compare to SNE that uses only one data set (not the user data), mSNE that uses both views and assumes a single set of common latent variables, and neighbourhood component analysis (NCA) that is a supervised method which assumes classes instead of another full-blown data view. We leave out comparison to MRE because code is not available, and the method would need some further development to be applicable for information visualisation; it is not obvious which latent variables to visualise.

We used three different data sets for comparison: scientific articles from [10], Reuters Corpus Volume 1 (RCV1), and Heart Disease data set [11] from the UCI repository [12]. We show numerical data for all but only have space to show the visualizations on one (RCV1). We use available class information as ground truth for user interest, and simulate additional structured noise to the user data (alternative classes and unstructured noise). For NCA we gave the advantage of using ground truth.

Performance is evaluated numerically by measuring the separability of the ground truth classes on the 2D visualization. The logic is that since we know the ground truth classes to be relevant in the sense that they inhabit the shared data space, and random errors are more likely to mix up the classes than to separate them, a better visualization separates the class distributions better. As a measure of separation we use the (leave-one-out) performance of a k-nearest neighbor classifier on the visualization.

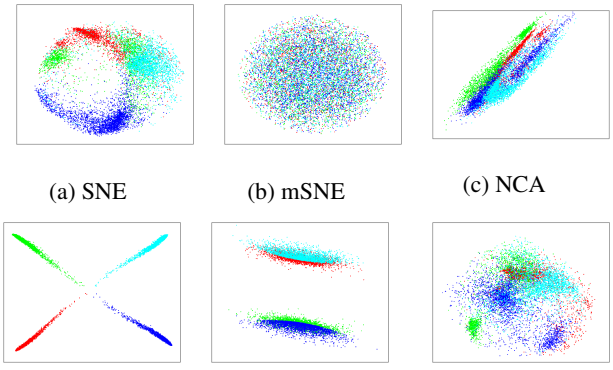
3.1. Reuters Corpus Volume 1

We used a subset of RCV1-v2 corpus, first used by [13]. The subset is a document-term matrix containing $N = 9,625$ documents which are divided into four categories, "C15", "ECAT", "GCAT", and "MCAT". For each document, feature vectors are generated by the standard TF-IDF weighting scheme. For details about the RCV1 corpus see [14].

Figure 2 shows our method finds the relevant structure clearly by inferring well-separated class-specific clusters. In this figure, categories are shown by colors of the dots: red, green, blue, and cyan represent "C15", "ECAT", "GCAT", and "MCAT", respectively. Some of the relevant structure is visible in the user-data-specific latent variables (Fig. 2e), and also in the irrelevant data (Fig. 2f), indicating that the dimensionality of the relevant structure is higher than two-dimensional and hence a higher-dimensional display would be needed to display all of the ground truth. For the alternative methods, based on the figure, we see that mSNE fails for the task, SNE infers two (noisy) clusters with incorrect classes and NCA captures a single cluster. For the other two data sets the behaviour of our proposed method was analogous and it separated the clusters well using the shared latent variables. Due to the lack of space we only show quantitative indicators on those two data sets, in Table 1, instead of visualizations.

The classes have become separated quantitatively as well, as shown with 5-nearest neighbor classification results in the second column of Table 1. We can see that the proposed multi-view latent variable model outperforms other techniques with a clear margin. It may seem striking that even the supervised NCA is clearly worse; the reason is that the data contain also irrelevant classes, and NCA is not better in

distinguishing the relevant and irrelevant ones. The table also contains the results for the two other data set where again our method is the best.



(a) SNE (b) mSNE (c) NCA

(d) our model (e) irrelevant user data (f) irrelevant data

Fig. 2: Comparison of methods in visualizing RCV1. (a): SNE; (b): mSNE; (c): NCA; (d) our model. Additionally, (e) shows user-data-specific visualization for our model, and (f) the irrelevant aspects of the primary data, extracted by our model.

Technique \ Data Set		RCV1	Scientific Article	Heart Disease
SNE		20.27 %	60.66 %	52.15 %
m-SNE		65.13 %	62.56 %	45.21 %
NCA		19.51 %	33.18 %	29.7 %
Proposed Method	(K = 6)	2.56%	9.9 %	46.20 %
	(K = 8)	**0.76 %**	**0.47 %**	22.11 %
	(K = 10)	6.85 %	11.37 %	**1.65 %**

Table 1: Quality of visualizations measured by separability of ground truth class distributions in the visualization, measured by a leave-one-out 5-NN classifier. The best result for each data set is marked in bold font. To evaluate sensitivity of our method to the number of components (of which $K-2$ are used to explain away irrelevant variation), results are shown for three different values of K.

3.2. Multi-labelling

To demonstrate that different aspects of the same data set can be visualized for different users, according to what is relevant to them, we simulate two users who are interested in different aspects and give different feedbacks (labellings). We used the Abalone data set from the UCI repository. This is a classification data set aimed at predicting the age of abalones from physical measurements. The number of rings inside the shell is a significant factor for determining the age of an abalone. We thresholded the numbers into three categories where the first group contains 3 to 9 rings, the second group contains 10 to 16 rings and finally the last group contains 17 to 23 rings. We left out ring numbers having less than 5 samples.This resulted in only 9 discarded records, which is negligible in the total number of 4177.

The first user is interested in the age and assigns similarities according to ring numbers. The second user is interested in sex and gives feedback according to the three categories M, F, and I. Figure 3 shows the visualization of the shared latent variables of our method for the two different labellings. In 3a the ages become separated to an extent, and in 3b the sexes. The visualizations are different, as they should be for two users interested in different aspects, both having support in the data.

(a) Age (b) Sex

Fig. 3: Visualization of the Abalone data for two users interested in different aspects of the data. (a) The colors correspond to different age groups (red: group 1; green: group 2, blue: group 3). (b) Cyan: M; magenta: F; black: I.

4. CONCLUSION

We have introduced 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. Unlike manifold embedding-based dimensionality reduction methods which have not been designed for compressing dimensionalities to two for visualization on display, our proposed method is able to visualize well in two dimensions, by explaining away the irrelevant data with additional latent variables.

In this paper, we successfully demonstrated and compared the proposed method on multiple static data sets, where the ground truth came from categorizations of the data. A main future goal is to use similar techniques in interactive visualization, where user interaction data will be measured all the time, and the visualization needs to react faster.

We did not consider model order selection in this paper. Our expectation is that standard probabilistic methods, in particular automatic relevance determination (ARD), could be used for determining the total number of latent variables. It may turn out to be more difficult to choose how many latent variables are relevant for the user, in case the relevant subspace cannot be adequately presented in 2D. Our current hypothesis is that group sparsity combined with ARD would be sufficient.

5. REFERENCES

[1] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[2] Joshua B Tenenbaum, Vin De Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[3] Sam T Roweis and Lawrence K Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

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

[5] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik, "Dimensionality reduction: A comparative review.," Tech. Rep. TiCC-TR 2009-005, Tilburg University, 2009.

[6] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[7] Bo Xie, Yang Mu, Dacheng Tao, and Kaiqi Huang, "m-sne: Multiview stochastic neighbor embedding," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, pp. 1088–1096, 2011.

[8] Roland Memisevic and Geoffrey E Hinton, "Multiple relational embedding," in *Advances in Neural Information Processing Systems*, 2004, pp. 913–920.

[9] Jaakko Peltonen and Samuel Kaski, "Generative modeling for maximizing precision and recall in information visualization," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 579–587.

[10] Jaakko Peltonen, Max Sandholm, and Samuel Kaski, "Information retrieval perspective to interactive data visualization," in *EuroVis-Short Papers*, 2013, pp. 49–53.

[11] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, pp. 304–310, 1989.

[12] M. Lichman, "UCI machine learning repository," 2013.

[13] Deng Cai and Xiaofei He, "Manifold adaptive experimental design for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, pp. 707–719, 2012.

[14] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li, "Rcv1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.