Exploratory Observation Machine (XOM) with Kullback-Leibler Divergence for Dimensionality Reduction and Visualization

Kerstin Bunte\textsuperscript{1,2}, Barbara Hammer\textsuperscript{3}, Thomas Villmann\textsuperscript{4}, Michael Biehl\textsuperscript{2} and Axel Wismüller\textsuperscript{1,*}

1- University of Rochester, Depts. of Radiology and Biomedical Engineering
601 Elmwood Avenue, Rochester, NY 14642-8648 - U.S.A.
2- University of Groningen - Johann Bernoulli Institute for Mathematics and Computer Science, 9700 AK Groningen - The Netherlands
3- Bielefeld University, CITEC
Universitätsstraße 23, 33615 Bielefeld - Germany
4- University of Applied Sciences Mittweida, Department of MPI
Technikumplatz 17, 09648 Mittweida - Germany

Abstract. We present an extension of the Exploratory Observation Machine (XOM) for structure-preserving dimensionality reduction. Based on minimizing the Kullback-Leibler divergence of neighborhood functions in data and image spaces, this Neighbor Embedding XOM (NE-XOM) creates a link between fast sequential online learning known from topology-preserving mappings and principled direct divergence optimization approaches. We quantitatively evaluate our method on real world data using multiple embedding quality measures. In this comparison, NE-XOM performs as a competitive trade-off between high embedding quality and low computational expense, which motivates its further use in real-world settings throughout science and engineering.

1 Introduction

Various dimension reduction techniques have been introduced based on different properties of the original data to be preserved. The spectrum ranges from linear projections of original data, such as in Principal Component Analysis (PCA) or classical Multidimensional Scaling (MDS) to a wide range of locally linear and non-linear approaches, such as Isomap, Locally Linear Embedding (LLE) [1], Local Linear Coordination (LLC), or charting. For a comprehensive recent review on nonlinear dimensionality reduction methods, we refer to [2].

Recently, a novel approach for topology-preserving learning has attracted attention for advanced data processing. The Exploratory Observation Machine (XOM) [3] computes graphical representations of high-dimensional observations by a strategy of self-organized model adaptation. Although simple and computationally efficient, XOM enjoys a surprising flexibility to simultaneously contribute to several different domains of advanced machine learning, scientific data analysis, and visualization, such as structure-preserving dimensionality reduction and data clustering [3]. Among a large number of different distance measures even including non-metric distances, it has been proposed in [4] to apply advanced divergence measures such as the Kullback-Leibler divergence and the Itakura-Saito distance within the XOM framework. This idea is in line with recent approaches to introduce alternative dissimilarity measures.

*These authors (K.B. and A.W.) contributed equally to this work.
for data processing, such as Sobolev-distances or kernel based dissimilarity measures [5, 2], approaches based on information theory using divergences for data processing, e.g. clustering [6, 7], dimension reduction with MDS, or Stochastic Neighbor Embedding (SNE) [8]. In this contribution, we derive a variant of XOM, called Neighbor Embedding XOM (NE-XOM) that builds upon the generalized Kullback-Leibler Divergence as a dissimilarity measure between the neighborhood distributions of high-dimensional data and low-dimensional image vectors. We will describe the XOM algorithm and its NE-XOM extension in section 2, discuss the embedding results on two benchmark data sets in section 3, and conclude in section 4.

2 The Exploratory Observation Machine (XOM)

We briefly review the Exploratory Observation Machine (XOM) algorithm. For details, we refer to the literature [9]. XOM maps a finite number of data points \( x^i \in \mathbb{R}^D \) in observation space \( \mathcal{X} \) to low dimensional data points \( y^i \in \mathbb{R}^d \) in the embedding space \( \mathcal{E} \). The assignment is \( x^i \rightarrow y^i \) and typically \( d \ll D \), e.g. \( d = 2 \) for visualization purpose. The embedding space \( \mathcal{E} \) is priorly equipped with a structure, given by a number of sampling vectors \( s^j \in \mathbb{R}^d \), which corresponds to the final structure according to which data are represented. Essentially, this is a generalization of the prototypes as included in the Self Organizing Map (SOM). Reasonable choices for the sampling vectors \( s^j \) are: the location on a regular lattice structure in \( \mathbb{R}^d \) just as in SOM, the location at discrete positions \( \mathbb{R}^d \) to represent a finite number of class centers, the sampling according to a mixture of Gaussians centered in \( \mathbb{R}^d \) to represent a finite number of clusters, or the uniform sampling in a region of \( \mathbb{R}^d \) to indicate that the visualization of the data should occupy the full projection space. Unlike SOM, XOM does not project the sampling vectors \( s^j \) (generalization of prototypes) to the data space, rather, it projects the data to the embedding space. Nevertheless, the sampling vectors define receptive fields by a decomposition into points mapped closest to the sampling vectors

\[
R_j = \{ x^i \mid d_\mathcal{E}(s^j, y^i) \text{ is minimum for } s^j \} \tag{1}
\]

where \( d_\mathcal{E} \) refers to the distance in the embedding space. An approximate back projection of the sampling vector can be defined as the best match input vector

\[
\Psi(s) = x^i \text{ where } d_\mathcal{E}(s, y^i) \text{ is minimum}. \tag{2}
\]

The images \( y^i \) are initialized randomly and adapted iteratively during the training triggered by the structure of the embedding space. All \( y^i \) are adapted into the direction of the current \( s^j \) according to the distances between the best match input \( \Psi(s^j) \) and their counterparts \( x^i \) in the observation space \( \mathcal{X} \). For a given sampling vector \( s^j \) the adaptation rule is given by:

\[
y^i := y^i - \eta h_\sigma(d_\mathcal{X}(\Psi(s^j), x^i)) \frac{\partial d_\mathcal{E}(s^j, y^i)}{\partial y^i}, \tag{3}
\]

where \( \eta > 0 \) denotes the learning rate, \( d_\mathcal{X} \) refers to the distance in the data space, e.g. the Euclidean distance and \( h_\sigma(t) = \exp(-t^2/2\sigma^2) \) with \( \sigma > 0 \) defines the neighborhood cooperation. In this way the projections \( y \) are arranged around the priorly chosen structure elements \( s \) such that image vectors are close to the
same sampling vector if their corresponding data points $x$ are neighbored in the data space. As the SOM, XOM in its original form does not possess a cost function. However as proposed in [9], a variation following Heskes [10] by setting the best match input data vector to the average

$$\Psi(s) = x^i$$

where $\sum_j h_\sigma(d_X(x^i,x^j))d_\varepsilon(s,y^j)$ is minimum

leads to the cost function:

$$E_{XOM} \sim \int \sum_i \delta_{\Psi(s),x^i} \cdot \sum_{j=1}^N h_\sigma(d_X(x^i,x^j)) \cdot d_\varepsilon(s,y^j) \cdot p(s)ds$$

where $\delta$ denotes the Kronecker delta.

### 2.1 XOM with generalized Kullback-Leibler Divergence

A recent and very powerful proposal for data visualization is SNE. It aims in finding projections such that the pairwise distribution of points in the data and embedding space are approximately the same measured by the Kullback-Leibler (KL) divergence. SNE has the drawback that, unlike in SOM or XOM, no prior structure of the projection space is involved, e.g. it is not intended to introduce a structuring component in the form of a lattice of sampling vectors. Like many other visualization techniques, SNE has a computational and memory complexity that is quadratic in the number of data points. The complexity of XOM can be easily controlled by the structure definition and is linear with the number of points and the number of sampling vectors. We propose to combine the ideas of XOM with the cost function as proposed by SNE.

Based on the cost function (5) we are able to define new learning rules for the XOM algorithm based on the generalized KL divergence for not normalized positive measures $p$ and $q$:

$$D_{GKL}(p \parallel q) = \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx - \int [p(x) - q(x)] dx$$

In contrast to [11], however, we do not use the KL divergence as a distance measure within the original or the embedding space, but as a dissimilarity measure between the two spaces. We define the cooperativity functions $h_\sigma(d_X(x^i,x^j))$ and $g_\gamma(d_\varepsilon(s,y^j))$ in the same way as shown in Sec. 2 to model the neighborhoods in the original space and embedding space. Based on these settings, we define a novel cost function using the divergence (6):

$$E_{GKL} \sim \int \sum_i \delta_{\Psi^{GKL}(s),x^i} \sum_j \left[ h_\sigma(d_X(x^i,x^j)) \ln \left( \frac{h_\sigma(d_X(x^i,x^j))}{g_\gamma(d_\varepsilon(s,y^j))} \right) \right] p(s)ds$$

where the best match data point for $s$ is defined as:

$$\Psi^{GKL}(s) = x^i$$

such that $\sum_j \left[ h_\sigma(d_X(x^i,x^j)) \ln \left( \frac{h_\sigma(d_X(x^i,x^j))}{g_\gamma(d_\varepsilon(s,y^j))} \right) \right] - h_\sigma(d_X(x^i,x^j)) + g_\gamma(d_\varepsilon(s,y^j))$ is minimum.

This leads to the online learning update rule for a given sampling vector $s$:

$$y^k = y^k - \frac{\eta}{\gamma_2} \left[ h_\sigma(d_X(\Psi^{GKL}(s),x^k)) - g_\gamma(d_\varepsilon(s,y^k)) \right] \frac{\partial d_\varepsilon(s,y^k)}{\partial y^k}.$$
emphasizes attraction and predominantly optimizes ‘continuity’, such that small distances in $\mathcal{X}$ lead to small distances in $\mathcal{E}$. The additional repulsive term in Eq. (9) is intended to facilitate optimization of ‘trustworthiness’, such that big distances in $\mathcal{X}$ enforce big distances in $\mathcal{E}$.

It is also possible to use this algorithm without a defined structure, one could simply change the definition of the sampling vectors, as inspired by [12, 13], in such a way that they are selected in close proximity to the image vector positions. Therefore, instead of choosing a sampling vector randomly out of a given distribution, we run through the images $\mathbf{y}^i$ and choose a sampling vector $s^i = \mathbf{\tilde{y}}^i$ drawn from a distribution centered around the actual images $\mathbf{y}^i$, e.g. from a Gaussian, a localized uniform, or a t-distribution. The algorithm, in the following called Neighbor Embedding XOM (NE-XOM) thus changes to:

**Step 1** - Compute pairwise distances $d_{\mathcal{X}}(x^i, x^j)$.

**Step 2** - Randomly initialize ‘image vectors’ $\mathbf{y}^i \in \mathcal{E}, i = 1, \ldots, N$ corresponding to each input vector $x^i$.

**Step 3** - Run through the randomized set $\mathbf{y}^i$, where one complete run is referred to as one epoch. For every $\mathbf{y}^i$, find a sampling vector drawn from a low variance distribution centered around $\mathbf{y}^i$. Subsequently, perform the update of all image vectors following Eq. (9). Another image vector is chosen and the procedure is repeated until a maximal number of epochs is reached. The final positions of the vectors $\mathbf{y}$ represent the output of the algorithm.

### 3 Experiments

In this section, we present results of NE-XOM on a real-world benchmark data set and quantitatively compare several widely used embedding techniques using multiple quantitative embedding quality measures as described in the literature, namely trustworthiness/continuity [14], Sammon’s stress [15], Spearman’s $\rho$ [16], and Pearson’s $r$ correlation.

#### 3.1 Wine

The wine data available at [17] contains 178 samples in 13 dimensions divided in three classes. As proposed in [18] we first transformed the data to have zero mean and unit variance features. NE-XOM was trained for $t_{\text{max}} = 50$ epochs with a learning rate annealing scheme $\eta(t) = \eta_1 \cdot \left( -\exp\left( \frac{\log(\eta_1)}{t_{\text{max}}} \right) \cdot t \right)$ with $\eta_1 = 0.1$ and $\eta_2 = 0.001$. The cooperativity functions $h_\sigma$ and $g_\gamma$ were chosen as Gaussians and their variance was annealed using the same scheme as for the learning rate with a local $\sigma_1$ value equal to the 80% percentile of the squared Euclidean distances for every point to all other points, $\sigma_2 = 4$. The prototypes were initialized with PCA in two dimensions. Various embedding quality measures can be found in table 1.

#### 3.2 USPS digits

The USPS dataset consists of images of hand written digits of a resolution of $16 \times 16$ pixel. We normalized the data to have zero mean and unit variance features, using the first 800 observations per class for the digits $\in \{0, 1, 2, 3, 4\}$, resulting in 4000 samples. The embedding obtained from NE-XOM learning with 50 epochs, globally annealed $\sigma$ for all data points and annealed $\gamma$ (same annealing scheme as used in wine data) is shown in Fig. 1. The parameters were chosen: $\eta_1 = 0.5$, $\eta_2 = 0.05$, $\gamma_1 = 0.5$, $\gamma_2 = 0.1$, $\sigma_1 = 25$ and $\sigma_2 = 4$. The prototypes were initialized using 2D PCA. Values for several quality measures for different methods are shown in Fig. 2 and Table 1. Interestingly, the embedding
As can be concluded from the results in both data sets, none of the compared algorithms excels with regard to all quality measures. NE-XOM outperforms slightly with regard to Spearman’s $\rho$ and Pearson’s $r$, SNE slightly in terms of trustworthiness and continuity. Interestingly, the embedding obtained from the NE-XOM show better continuity/trustworthiness values than widely used methods like Isomap, LLE, or Sammon’s mapping. The continuity of LLE is better on this USPS data set, but its visualization is poor, because all points are collapsed on a line. Sammon’s stress is only outperformed by Sammon’s mapping itself in this data set, and with Spearman’s $\rho$ and Pearson’s $r$, the best values can be obtained with NE-XOM compared to traditional methods.

4 Conclusions and Outlook

In this contribution, we have introduced an extension of the Exploratory Observation Machine (XOM) for structure-preserving dimensionality reduction. Based
Table 1: Summary for the benchmark data sets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Wine Sam. St.</th>
<th>Wine Sp. rho</th>
<th>Wine Pear. r</th>
<th>USPS Sam. St.</th>
<th>USPS Sp. rho</th>
<th>USPS Pear. r</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE-XOM</td>
<td>0.07</td>
<td>0.89</td>
<td>0.88</td>
<td>0.14</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>SNE</td>
<td>0.12</td>
<td>0.78</td>
<td>0.74</td>
<td>0.16</td>
<td>0.52</td>
<td>0.54</td>
</tr>
<tr>
<td>Sammon</td>
<td>0.07</td>
<td>0.87</td>
<td>0.86</td>
<td>0.12</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>LLE</td>
<td>0.17</td>
<td>0.66</td>
<td>0.64</td>
<td>0.27</td>
<td>0.29</td>
<td>0.35</td>
</tr>
<tr>
<td>Isomap</td>
<td>0.18</td>
<td>0.80</td>
<td>0.77</td>
<td>0.43</td>
<td>0.29</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Acknowledgements

This work was supported by the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)" under project code 612.066.620 and the University of Rochester, Deps. of Radiology and Biomedical Engineering.

References