

# Linear basis-function t-SNE for fast nonlinear dimensionality reduction

A. Gisbrecht, B. Mokbel, and B. Hammer,

University of Bielefeld,  
CITEC Center of Excellence,  
Universitätsstrasse 21-23,  
33615 Bielefeld, Germany

Email: {agisbrec|bmokbel|bhammer}@techfak.uni-bielefeld.de

**Abstract**—t-distributed stochastic neighbor embedding (t-SNE) constitutes a nonlinear dimensionality reduction technique which is particularly suited to visualize high dimensional data sets with intrinsic nonlinear structures. A major drawback, however, consists in its squared complexity which makes the technique infeasible for large data sets or online application in an interactive framework. In addition, since the technique is non parametric, it possesses no direct method to extend the technique to novel data points. In this contribution, we propose an extension of t-SNE to an explicit mapping. In the limit, it reduces to standard non-parametric t-SNE, while offering a feasible nonlinear embedding function for other parameter choices. We evaluate the performance of the technique when trained on a small subpart of the given data only. It turns out that its generalization ability is good when evaluated with the standard quality curve. Further, in many cases, it obtains a quality which approximates the quality of t-SNE when trained on the full data set, albeit only 10% of the data are used for training. This opens the way towards efficient nonlinear dimensionality reduction techniques as required in interactive settings.

## I. INTRODUCTION

Electronic data sets available today increase dramatically in size and complexity. Often, it is no longer possible to directly infer relevant information from the data, to directly inspect the data by hand, or even to formalize a clear learning objective in a given setting. This turns visualization techniques to more and more relevant tools which provide a feasible interface to the data. This way, humans can easily browse through massive volumes of data sets relying on their astonishing cognitive capabilities for visual perception. Based on this first impression, a refinement of what information could be of interest can be done by humans, eventually resulting in a clearly defined machine learning task, for example. In this realm, the field of scalable visual analytics has become a rapidly emerging research area which deals with methods to automatize and evaluate visualization techniques connected with intelligent inference and machine learning techniques in large systems and to link these techniques towards human users and their feedback [12].

One important technology in this context is given by methods which project large high dimensional data sets onto the plane such that as much information as possible is preserved. Such dimensionality reduction techniques offer a basic functionality to visually large data volumes to humans by means

of a collection of points in the plane where, ideally, relevant structural aspects of the data such as outliers, cluster structures, etc. can directly be captured. Further, assuming the availability of additional information such as class labels, the relation of the inherent data structure to this auxiliary information can easily be displayed by means of an appropriate coloring. In the last years, many dimensionality reduction techniques have been developed in this context, see e.g. [4], [14], [22]. These techniques enclose more and more sophisticated nonlinear dimensionality reduction methods such as t-distributed stochastic neighbor embedding, locally linear embedding, maximum variance unfolding, multiple relational embedding [11], [22], [10], [2], [17], [19], [24], [23], and the like, which display a huge flexibility to locally emphasize important structural aspects such as an underlying low dimensional nonlinear manifold or cluster structures.

Despite of this fact, however, the techniques which are usually used by applicants outside the field reduce to simple principle component analysis (PCA) or multi-dimensional scaling, see e.g. [3] for a recent discussion resulting from a meeting of experts in the field. These techniques are essentially linear (assuming classical metric multidimensional scaling), such that their flexibility to display local nonlinear structures on the plane is strictly limited. What are the reasons that, still, these techniques are often preferred in real life applications, neglecting recent nonlinear projection techniques? At present, all methods are publicly available.<sup>1</sup> Further, techniques such as t-SNE do not require expert knowledge to appropriately choose the model parameters - in fact, t-SNE comes with a single parameter, the so-called perplexity, which can usually be set to a default value. Thus, availability is not an issue. However, one of the most severe problems of t-SNE in comparison to PCA is given by the fact that t-SNE scales quadratic with the number of input data. In addition, t-SNE being non-parametric, it provides a mapping of the given points only, making it necessary to rerun t-SNE as soon as a novel data point should be mapped. While t-SNE runs in reasonable time for a few hundred data points, it can no longer be used in an

<sup>1</sup>The dimensionality reduction toolbox of van der Maaten, for example, provides an easy to use interface to most currently available techniques as well as a good documentation, see [http://homepage.tudelft.nl/19j49/Matlab\\_Toolbox\\_for\\_Dimensionality\\_Reduction.html](http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html)

interactive fashion for a few thousand data points, and it is not applicable at all for larger data sets, simply because of the required training time, assuming standard desktop computers. Thus, the squared complexity of t-SNE and similar techniques currently constitutes one of the most severe problems, because of which it cannot yet be used in interactive scenarios where an immediate response is required and more than just a few hundred data points should be inspected. PCA, on the contrary, constitutes a simple linear time technique which provides an explicit linear embedding of the data in low dimensions.

In this contribution, we investigate the possibility to speed up t-SNE by means of an explicit embedding trained using t-SNE. That means, instead of the standard non-parametric mapping provided by t-SNE, a parametric embedding in the form of a generalized linear model based on kernels becomes available to project the data points. We particularly emphasize on the generalization ability of this approach. Using typical benchmark scenarios, we investigate whether a linear basis-function t-SNE mapping trained on a fraction of all data only provides a good visualization for the full data set. If this was the case, an immediate speed-up technique which makes t-SNE useful in interactive settings would result. Further, it would open the way towards projection scenarios where data are not given a priori, but online mapping of additional points is required.

Several approaches have been proposed in the literature to extend non-parametric nonlinear dimensionality reduction techniques to explicit mappings. The approach [1] proposes interpolation techniques for multidimensional scaling and the generative topographic mapping. The contribution [20] considers visualization by means of a kernel function optimized according to a quadratic cost function. In the approach [21] t-SNE is combined with deep feedforward networks to arrive at a highly nonlinear embedding function. Recently, a variety of well-known non-parametric nonlinear dimensionality reduction techniques has been put into a general framework based on the notion of cost functions [4]. Based on this formulation, an extension of the techniques to parametric mappings by means of optimization has been proposed. Albeit these techniques provide promising results, up to our knowledge, no approach investigates the generalization ability of the mappings if trained on a very small subset of the data only, and no approach comes with a very fast training method for the parametric mapping which can be used in interactive frameworks.

In this contribution, we extend t-SNE to parametric mappings (linear basis-function t-SNE) which, in the limit of small bandwidth, exactly resembles t-SNE. Hence, unlike globally linear techniques such as PCA, it has a huge flexibility to locally emphasize relevant structures of the data. We propose different ways how to train the parameters of the mapping. It turns out that a particularly fast and simple technique by explicit linear regression on a training set delivers very good results. Then, we investigate the generalization capability of linear basis-function t-SNE in several benchmark scenarios. For this purpose, we compare the quality of the mapping on

the training set, the test set, and the quality of a full non-parametric t-SNE embedding given all points. It turns out that the quality of the test set approximates the quality of the training set. Further, in many cases, the quality of the linear basis-function t-SNE embedding approximates the quality of non-parametric t-SNE when trained for the full data set. This property, however, severely depends on the characteristic of the given data set. In some cases, it is not yet possible to infer important structural forms from a subset of the given data only, necessarily resulting in a lower quality when training t-SNE with a subset only.

For the evaluation of the quality of dimensionality reduction, we use the co-ranking framework as proposed by Lee and Verleysen [15], [16]. In this framework, which is very similar to the evaluation measures as proposed in [23], Lee and Verleysen consider neighborhoods of the  $k$  nearest points in the original space and the projection space. The quality sums up over the size of these intersections, normalized in an appropriate way. This way, a quality measure results for every  $k$  which ranges from 0 to 1, 1 indicating that the  $k$ -neighborhoods of points are perfectly preserved when projecting the data. This evaluation measure, however, encounters a problem: it is not insensitive with respect to the overall number of data used for the mapping and its evaluation. Further, like nonlinear embedding itself, it is of quadratic complexity. Here, we propose an approximation mechanism based on sampling which makes the technique insensitive to the overall number of data and which speeds up the evaluation. Since a quantitative evaluation measure can serve as a way to automatically set meta-parameters of a dimensionality reduction technique (such as e.g. the choice of the appropriate dimensionality reduction technique itself), this approximation constitutes another important step towards a fast interactive method.

## II. T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

t-distributed stochastic neighbor embedding (t-SNE) has been proposed in [22] as a highly flexible dimensionality reduction (DR) technique. It is based on a probabilistic modeling of points in the original space and the projection space. It tries to preserve these probabilities as much as possible as measured by the Kullback-Leibler distance. Assume data points  $\mathbf{x}_i \in \mathbb{R}^n$  are given. The Euclidean distance induces pairwise probabilities  $p_{ij} = \frac{1}{2}(p_{i|j} + p_{j|i})$ , where

$$p_{i|j} := \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2 / 2\sigma_i^2)}.$$

Every data point is mapped to a projection point  $\mathbf{y}_i \in \mathbb{R}^d$ , where, typically  $d = 2$ . These projections also induce pairwise probabilities

$$q_{ij} := \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}.$$

Unlike the probabilities in the original space, the Student's t-distribution is used in the projection space. Because of its long tails, this allows to match medium sized distances in the original space with large distances in the projection space,

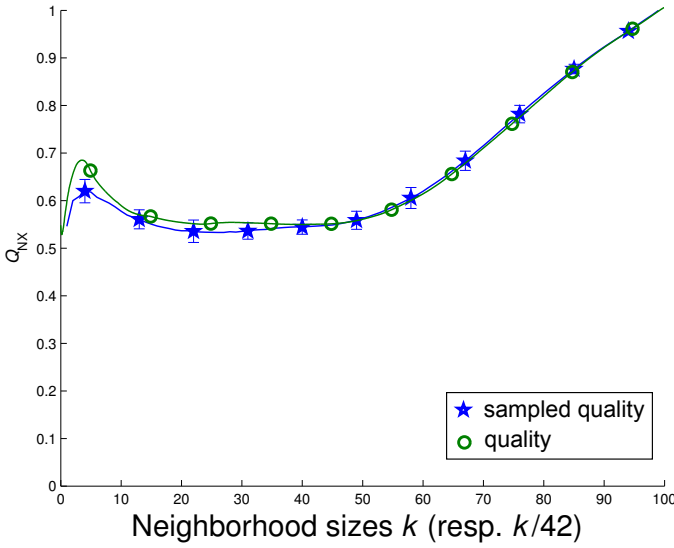


Fig. 1. Quality graph for chromosomes for all 4,200 data compared to a quality curve obtained by sampling.

thus preventing the often encountered crowding problem when projecting data to low dimensionality: in low dimensions, on average, points are necessarily further away than in high dimensionality due to the limited space. Another more subtle discussion of phenomena such as norm concentration in high dimensions and its consequences for dimensionality reduction has recently been published in [25]. t-SNE aims at finding projections  $\mathbf{y}_i$  such that the difference of these probabilities as measured by the Kullback-Leibler divergence

$$\text{KL}(P||Q) := \sum_i \sum_j p_{ij} \log \left( \frac{p_{ij}}{q_{ij}} \right)$$

is minimized. Typically a gradient technique is used for optimization. The parameter  $\sigma_i$  is determined based on the data set such that the effective number of neighbors equals a fixed meta-parameter of the algorithm, the perplexity. Usually, a default value for the perplexity (such as 15) is appropriate, and the technique is not very sensitive with respect to this parameter.

t-SNE yields a projection of points only, and does not provide an explicit embedding mapping. It has been discussed e.g. in [4], how the technique can be extended to out-of-sample points by means of a minimization of the cost terms for novel entries. Still, no explicit mapping is given this way. Here, we consider a parameterized mapping on top of t-SNE given by an explicit embedding.

### III. LINEAR BASIS-FUNCTION T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

We consider a generalized linear mapping which has the following form

$$\mathbf{x} \mapsto \mathbf{y}(\mathbf{x}) = \sum_l \alpha_l k(\mathbf{x}_l, \mathbf{x})$$

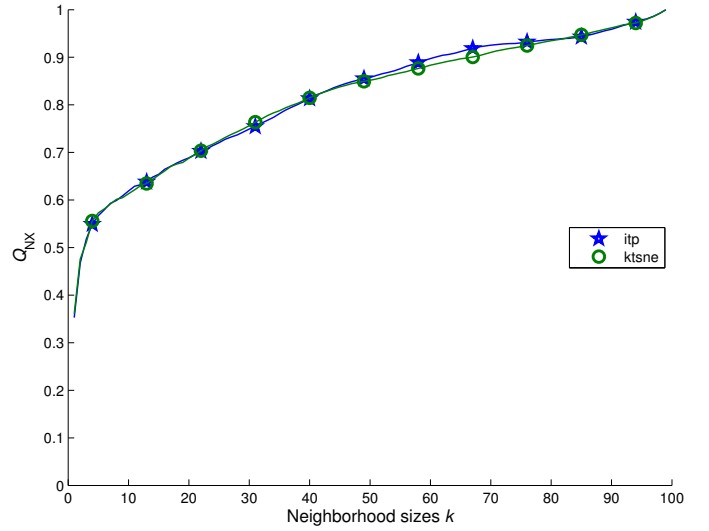


Fig. 2. Results on the chromosomes data set for a linear basis-function t-SNE when trained for 10% of the data. The result on the test set is depicted. We show the result when training takes place by optimizing the t-SNE cost function with the kernel mapping (ktsne) versus a direct interpolation (itp) of the training set mapped using t-SNE.

where  $\alpha_l \in \mathbb{R}^d$  and  $\{\mathbf{x}_l | l\}$  is a fixed sample of data points.  $k$  is an appropriate basis function such as the Gaussian kernel or the recently proposed ‘nearly’ parameterless ELM kernel [7]. We also consider the corresponding normalized variant

$$\mathbf{x} \mapsto \mathbf{y}(\mathbf{x}) = \frac{\sum_l \alpha_l k(\mathbf{x}_l, \mathbf{x})}{\sum_l k(\mathbf{x}_l, \mathbf{x})}$$

which is beneficial for geometric mappings since it guarantees to map data points into the convex hull of the images of the support vectors  $\mathbf{x}_l$ .

The parameters  $\alpha_l$  of the mapping have to be determined based on a given training sample  $\{\mathbf{x}_l | l\}$  of points. There exist different possibilities to determine these parameters:

- 1) t-SNE cost function optimization of the parameters: assume a sample of points  $\{\mathbf{x}_l | l\}$  is given. We can insert these points and their projections  $\mathbf{y}_l = \mathbf{y}(\mathbf{x}_l)$  in parameterized form in the cost function of t-SNE resulting in the costs  $E := \text{KL}(P||Q)$  where  $Q$  is determined based on the projections  $\mathbf{y}(\mathbf{x}_l)$  provided by the kernel mapping. Optimization can be performed by gradient techniques where the gradient of  $E$  with respect to  $\alpha_l$  is

$$4 \sum_i \sum_j (p_{ij} - q_{ij}) (1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1} (\mathbf{y}_i - \mathbf{y}_j) k(\mathbf{x}_i, \mathbf{x}_j)$$

for the standard kernel. This way, parameters which optimize the underlying t-SNE cost function can be determined.

- 2) Interpolation of a given set of points: Alternatively, we can turn the problem into a standard machine learning task. For a given set of points  $\{\mathbf{x}_l | l\}$ , we determine projection points  $\mathbf{y}_l$  such that these points minimize

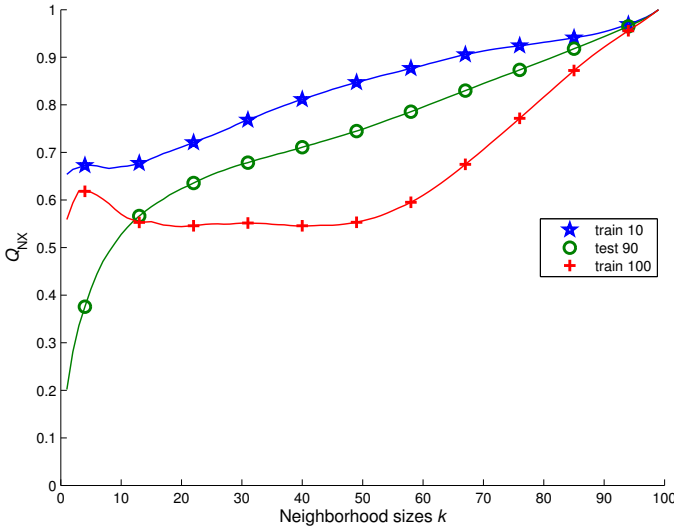


Fig. 3. Results on the chromosomes data set when evaluating the quality for the full data set, t-SNE trained on 10%, and the resulting kernel mapping for the remaining 50% of the data.

the objective of t-SNE. Afterwards, we learn the parameters of the mapping  $\mathbf{x} \rightarrow \mathbf{y}(\mathbf{x})$  by treating the set of pairs  $\{(\mathbf{x}_l, \mathbf{y}_l) \mid l\}$  as learning set for a classical regression problem. We can treat this regression problem as generalized linear regression which can be solved algebraically taking the pseudo-inverse. This way, the mean squared error is minimized on the training set. We obtain the matrix  $\mathbf{A}$  of parameters via  $\mathbf{A} = \mathbf{K}^{-1}\mathbf{Y}$  where  $\mathbf{K}$  is the Gram matrix with entries  $k(\mathbf{x}_i, \mathbf{x}_j)$  or the normalized variant thereof, respectively,  $\mathbf{Y}$  denotes the matrix of projections  $\mathbf{y}_l$ , and  $\mathbf{K}^{-1}$  refers to the pseudo-inverse.

Both methods are cubic in the number of points used for training. However, our expectation is that it is sufficient to train the mapping on a small subset of the given data set only, and to extrapolate to all other points afterwards. This way, the squared complexity required to train t-SNE for all points is reduced to an only linear one. This claim will be tested in the experiments.

#### IV. EVALUATION OF DIMENSIONALITY REDUCTION

In [15], [16], a framework to quantitatively evaluate dimensionality reduction has been proposed, which summarizes a number of previous approaches in a unifying co-ranking framework. The basic ideas are as follows: for every point  $\mathbf{x}_l$  and its projection  $\mathbf{y}_l$ , the indices of the  $k$  nearest neighbors are determined in the original data space  $N_k^i(\mathbf{x}_l)$  and projection space  $N_k^o(\mathbf{y}_l)$ . The quality for the parameter  $k$  is the quantity

$$q_k = \frac{1}{kN} \sum_l |N_k^i(\mathbf{x}_l) \cap N_k^o(\mathbf{y}_l)|$$

where  $N$  denotes the overall number of points. A value in  $[0, 1]$  results with 1 indicating that the  $k$  neighbors in the projection space and the original data space are identical.

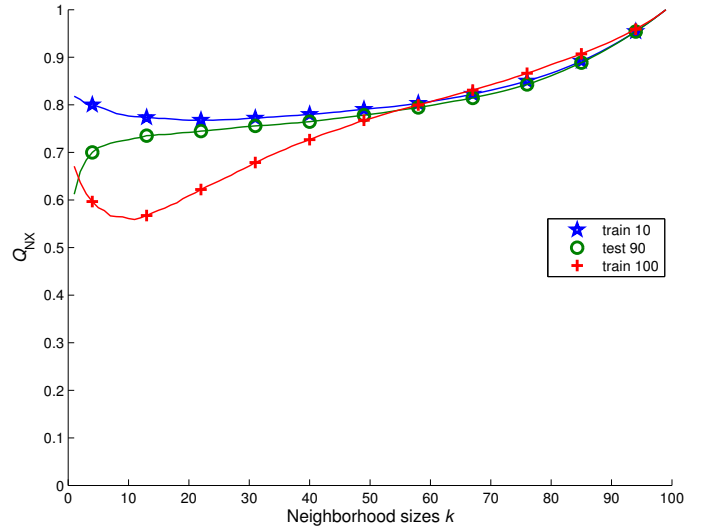


Fig. 4. Quality curves for the sphere data set.

This framework results in a quality curve which evaluates the quality in how far neighborhoods are preserved for a neighborhood size depending on  $k$ .

The quality curves provide a state of the art evaluation scheme to formally compare data projections. However, the framework has a severe drawback: it has quadratic complexity with respect to the number of points. Thus, it is infeasible for large data sets, and it can hardly be used in an interactive setting to evaluate dimensionality reductions on demand. Another drawback of the quality framework is that it severely depends on the number of data points. Assume numbers  $N_1$  and  $N_2$  of data are sampled from the same manifold, then it is generally not possible to compare the curves resulting for these two sets of points of different size because  $q_{N_1}(k) \neq q_{N_2}(k)$  in general.

To evaluate dimensionality reduction techniques in reasonable time and to compare dimensionality reduction performed using sets of different size, we propose a simple sampling strategy: for all settings, a fixed set of points is sampled, and the quality is evaluated. This is repeated 10 times and the mean value is taken. We depict the result of such a quality curve together with the standard deviation in comparison to the original quality curve in Fig. 1 for one example data set (the Chromosomes data set which consists of 4,200 points). Here we sample 10% of the data. Due to the ratio of sampled points, we expect that the quality graphs are related by means of the equation  $q_N(k) = q_{N/10}(k/10)$ . It can be observed that a deviation exists for small values while there is very good agreement already for a neighborhood size 10. It can be expected that the resulting curves can be used to compare the quality of different projections if the size of the sampled points coincides. This way, quality evaluation is linear, assuming fixed sample size, and different sized mappings can be compared.

#### V. EXPERIMENTS

We use the following data sets:

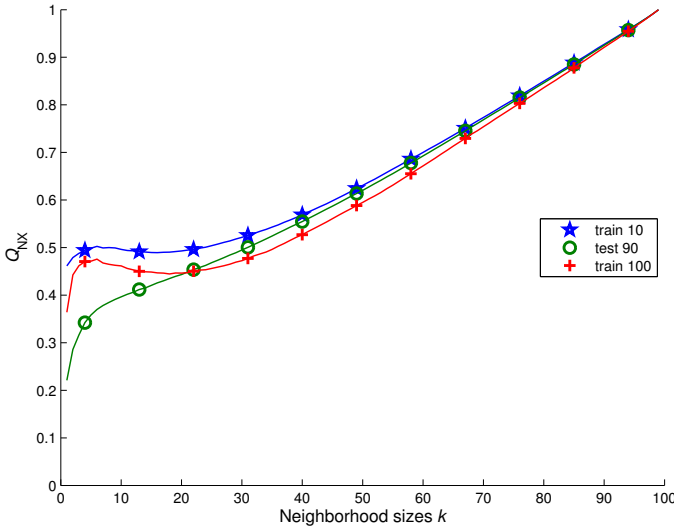


Fig. 5. Quality curves for the mnist data set.

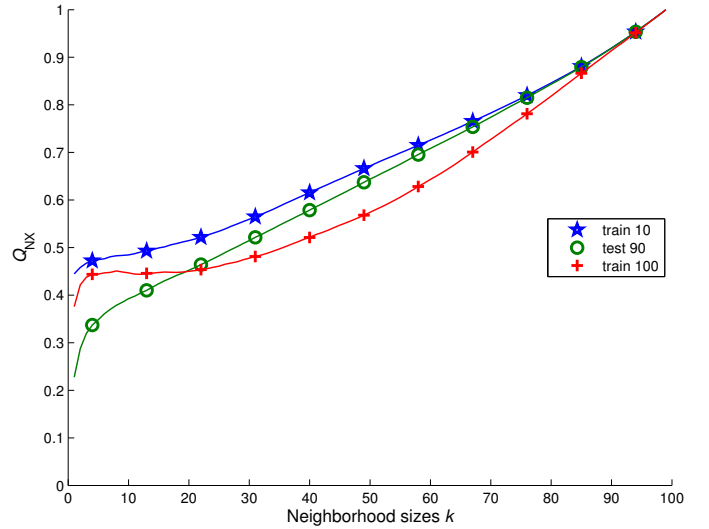


Fig. 6. Quality curves for the usps data set.

- 1) Chromosomes: The Chromosomes data constitute a benchmark from cytogenetics. 4,200 human chromosomes from 21 classes (the autosomal chromosomes) are represented by grey-valued images. These are transferred to strings measuring the thickness of their silhouettes. These strings are compared using edit distance with insertion/deletion costs 4.5 [18].
- 2) Sphere: The sphere data consists of 10,000 data points on a sphere in three dimensions with radius 1, data are colored according to their  $z$  coordinate [14].
- 3) Mnist: This data set consists of 60,000 points with 768 dimensions representing the ten digits. We use a subsample of 10,000 points [13].
- 4) Usps: The usps data set consists of 11,000 points with 256 dimensions also representing the digits from 0 to 9 taken from the US postal service [9].
- 5) Henon: The henon data set consists of 5,000 data points in three dimensions forming a chaotic attractor as described by the corresponding dynamical equations which constitute a generalization of the popular henon attractor to three dimensions [8].

For all settings, we use a normalized Gaussian kernel. We set the bandwidth  $\beta_i$  based on the characteristics of the data. It is set as the distance between  $\mathbf{x}_i$  and its first neighbor, scaled with 0.1 for chromosome, sphere, and henon data sets and 0.005 for usps and mnist. We evaluate all data sets taking quality curves based on 100 random samples with 10 repeats. For all settings, we compare the result of t-SNE when trained on the full data set versus a linear basis-function t-SNE mapping trained on randomly chosen 10% of the data only and extrapolated to the remaining 90%.

First, we evaluate the difference of the two different training techniques, an optimization of kernel parameters by means of an optimization of the t-SNE cost functions versus a direct interpolation of the data from a given training set obtained by t-SNE. The result is exemplarily shown in Fig. 2 for

the chromosomes data set. Obviously, no difference of the result can be observed. Therefore, we simply use the direct algebraic solution obtained by interpolation in the following. This method has the advantage that it can be directly used on top of alternative dimensionality reduction methods instead of t-SNE without any further adaptation.

Next, we test the ability of the kernel mapping to generalize by comparing the curves obtained when training t-SNE on 10% of the data set to the quality curves when extrapolating to the remaining 90% of the data set using a kernel mapping. We also evaluate the overall quality of linear basis-function t-SNE by comparing to the result obtained by t-SNE for the full data set. All quality curves are obtained by means of subsampling.

The results obtained for the chromosomes data are depicted in Fig. 3 and 8. Coloring of the projected data is according to the priorly known classes of the chromosomes as given in the data set. Interestingly, the quality when using t-SNE on 10% of the data only is better as compared to t-SNE when training on the full data set. The extrapolation based on kernel mapping is better than the explicit t-SNE mapping starting from neighborhood sizes  $k \approx 12$ , approaching the quality of the mapping for the training set for large  $k$ . The dimensionality reduction mapping as displayed in Fig. 8 clearly shows the characteristics of the kernel mapping. Unlike t-SNE for the full data set, priorly known clusters possess less margin in the projection when using only a subset for training because they are not yet apparent in this small subset. Nevertheless, all clusters clearly arrange at separate positions in the projection and this property is preserved in the extrapolation using the kernel mapping. When referring to the overall quality of the mapping as evaluated by the quality measure proposed by Lee and Verleysen, the latter mappings seem to better respect neighborhood structures as present in the data.

For the sphere data set, the result is similar. The quality curve is displayed in Fig. 4, clearly indicating that the kernel mapping gives better results already for small neighborhood



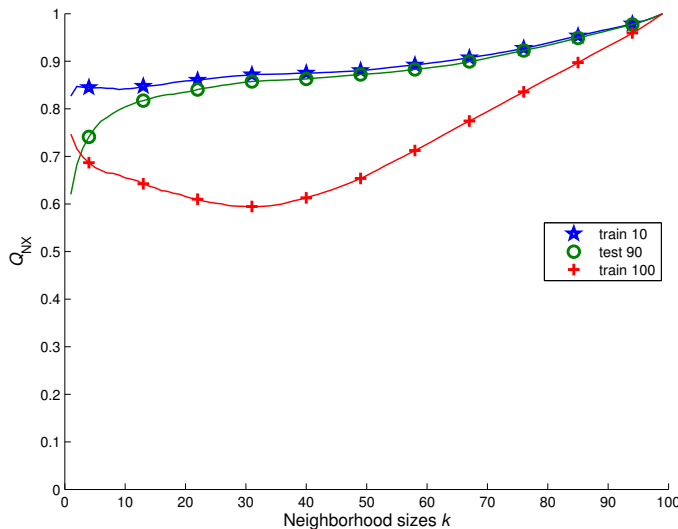


Fig. 7. Quality curves for the henon data set.

size as compared to full t-SNE. The projections of the data are depicted in Fig 9, whereby coloring is according to the z-axis of points in three dimensions. Again, the kernel mapping shows a very good generalization behavior as compared to the test set.

For mnist, the three quality curves are very close, as depicted in Fig. 5. Local clusters have less margin when using only 10% for training, but the generalization ability of the kernel mapping seems very good, as can be seen in Fig. 10. As similar behavior can be observed for the usps data set, see Fig. 6 and Fig. 11, and the henon data, see Fig. 7.

## VI. DISCUSSION

We have proposed a method to obtain an explicit nonlinear kernel mapping for dimensionality reduction based on t-SNE. We have extensively evaluated the generalization ability of this technique by training on a very small subpart of the given data only and extrapolating to the rest. Thereby, we restricted to random sampling assuming i.i.d. data such that the underlying structure is still observable. The effect of biased sampling or even active data selection will be the subject of future work. It turns out that, for random sampling, the generalization ability is acceptable in general, and that the resulting mappings are even better if evaluated in the co-ranking framework for medium sized and large values of the neighborhood. However, cluster structures are less apparent due to a smaller margin of the classes when training for a small subset of the data only. In most cases, still a first inspection of the data is possible based on the result obtained for a small subpart only, such that this techniques constitutes a promising approach if fast interactive techniques are required. By referring to an explicit functional form, linear basis-function t-SNE opens the way towards an easy integration of prior knowledge such as invariances which can directly be integrated into the kernel mapping. This constitutes the subject of future research.

**Funding:** This work has been supported by the German Res. Fund. (DFG), HA2719/4-1 and HA2719/7-1 and by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.

## REFERENCES

- [1] Bae, S.-H., Choi, J. Y., Qiu, J., and Fox, G. C. (2010). Dimension reduction and visualization of large high-dimensional data via interpolation. In *Proceedings of HPDC*, 10:203–214.
- [2] Belkin, M., Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–15396.
- [3] Michael Biehl, Barbara Hammer, Erzsébet Merényi, Alessandro Sperduti, Thomas Villmann: Learning in the context of very high dimensional data (Dagstuhl Seminar 11341). Dagstuhl Reports 1(8): 67-95 (2011)
- [4] K. Bunte, M. Biehl, B. Hammer (2012). A general framework for dimensionality reducing data visualization mapping. *Neural Computation*, in press.
- [5] Chih-Chung Chang, Chih-Jen Lin (2001), *LIBSVM: a library for support vector machines*, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [6] A. Frank, A. Asuncion (2010), UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, available at <http://archive.ics.uci.edu/ml/>
- [7] Frénay B, Verleysen M (2011). Parameter-insensitive kernel in extreme learning for non-linear support vector regression. *Neurocomputing* 74(16):2526-2531.
- [8] Gonchenko, S.V., Ovsyannikov, I.I., Simo, C., Turaev, D. (2005), Three-dimensional Henon-like maps and wild Lorenz-like attractors, Technical Report MP-ARC-2005-111, University of Texas.
- [9] Hastie, Tibshirani and Friedman, The elements of statistical learning, Springer, 2001.
- [10] He, X., Cai, D., Yan, S., and Zhang, H.-J. (2005). Neighborhood preserving embedding. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1208–1213 Vol. 2.
- [11] Hinton, G. and Roweis, S. (2003). Stochastic neighbor embedding. In *Advances in NIPS 15*, pages 833–840. MIT Press.
- [12] Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). Visual analytics: Scope and challenges. In Simoff, S., Boehlen, M. H., and Mazeika, A., editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer. LNCS.
- [13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998
- [14] Lee, J. A., Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer, 1st edition.
- [15] Lee, J. A., Verleysen, M. (2010). Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters* 31(14): 2248–2257.
- [16] Lee, J. A., Verleysen, M. (2009). Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomput.*, 72(7-9):1431–1443.
- [17] Memisevic, R. and Hinton, G. (2005). Multiple relational embedding. *Advances in NIPS*, 17:913–920.
- [18] Neuhaus, M., Bunke, H. (2006). Edit distance based kernel functions for structural pattern classification. *Pattern Recognition*, 39(10):1852–1863.
- [19] Roweis, S. T., Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326.
- [20] Suykens, J. A. K. (2008). Data visualization and dimensionality reduction using kernel maps with a reference point. *Neural Networks, IEEE Transactions on*, 19(9):1501–1517.
- [21] van der Maaten, L. J. P. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of AI-STATS*, 5:384–391.
- [22] van der Maaten, L. J. P. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.
- [23] Venna, J., Peltonen, J., Nybo, K., Aidos, H., and Kaski, S. (2010b). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR*, 11:451–490.
- [24] Weinberger, K. Q. and Saul, L. K. (2006). An introduction to nonlinear dimensionality reduction by maximum variance unfolding. *Proceedings of the 21st National Conference on Artificial Intelligence*.
- [25] Lee, J. A. and Verleysen, M. (2011) Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Proceedings of ICCS 2011*, 4:538–547.

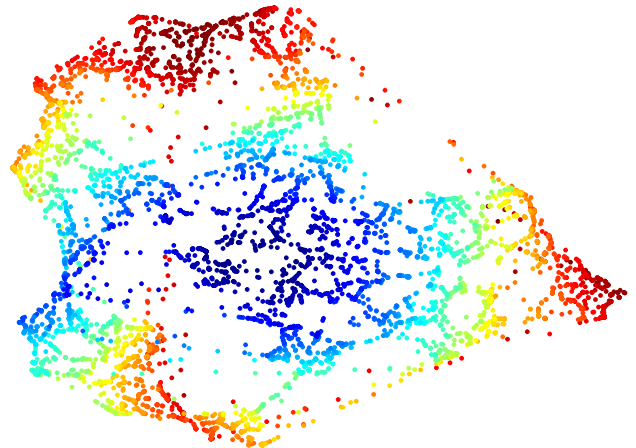
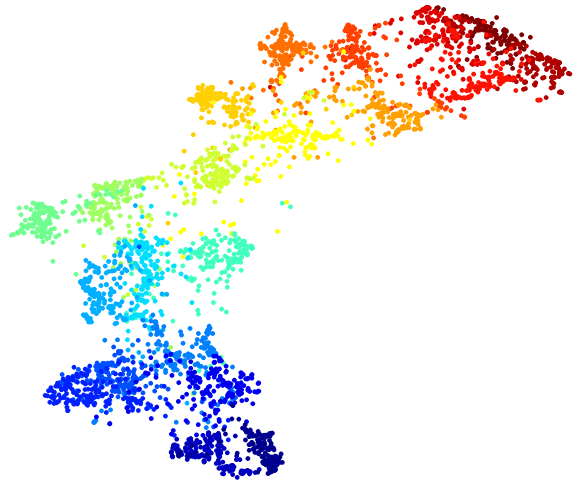
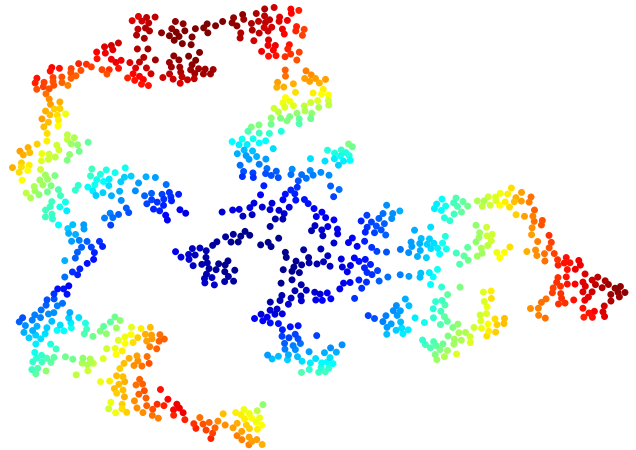
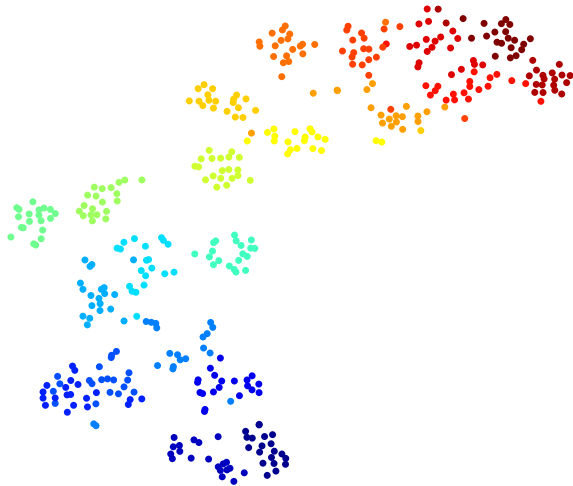
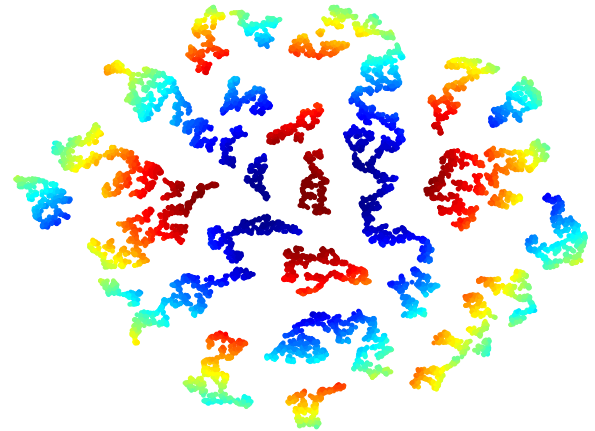
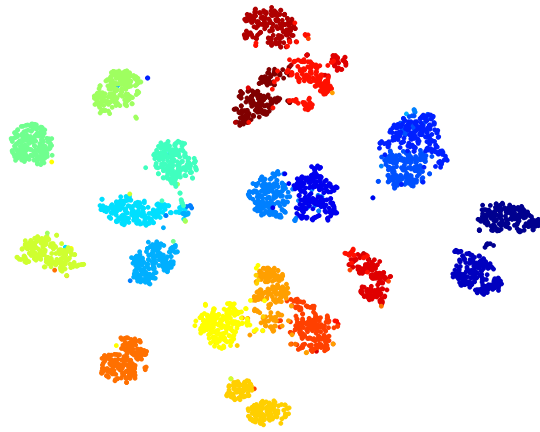


Fig. 8. Results on the chromosomes data set. From top to bottom: projection when using t-SNE for the full data set; projection when training t-SNE on 10% of the data set; projection when extrapolating for the remaining 90% using a kernel mapping

Fig. 9. Results on the sphere data set. From top to bottom: projection when using t-SNE for the full data set; projection when training t-SNE on 10% of the data set; projection when extrapolating for the remaining 90% using a kernel mapping

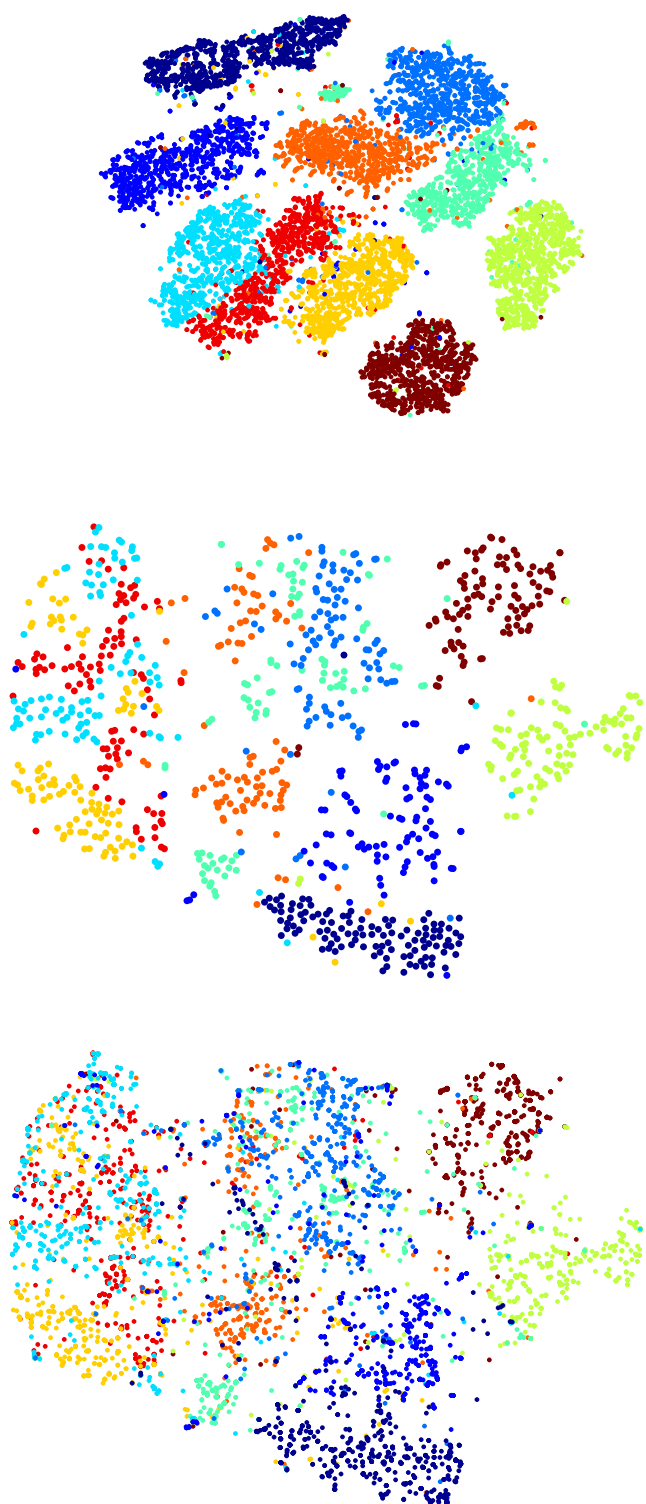


Fig. 10. Results on the mnist data set. From top to bottom: projection when using t-SNE for the full data set; projection when training t-SNE on 10% of the data set; projection when extrapolating for the remaining 90% using a kernel mapping

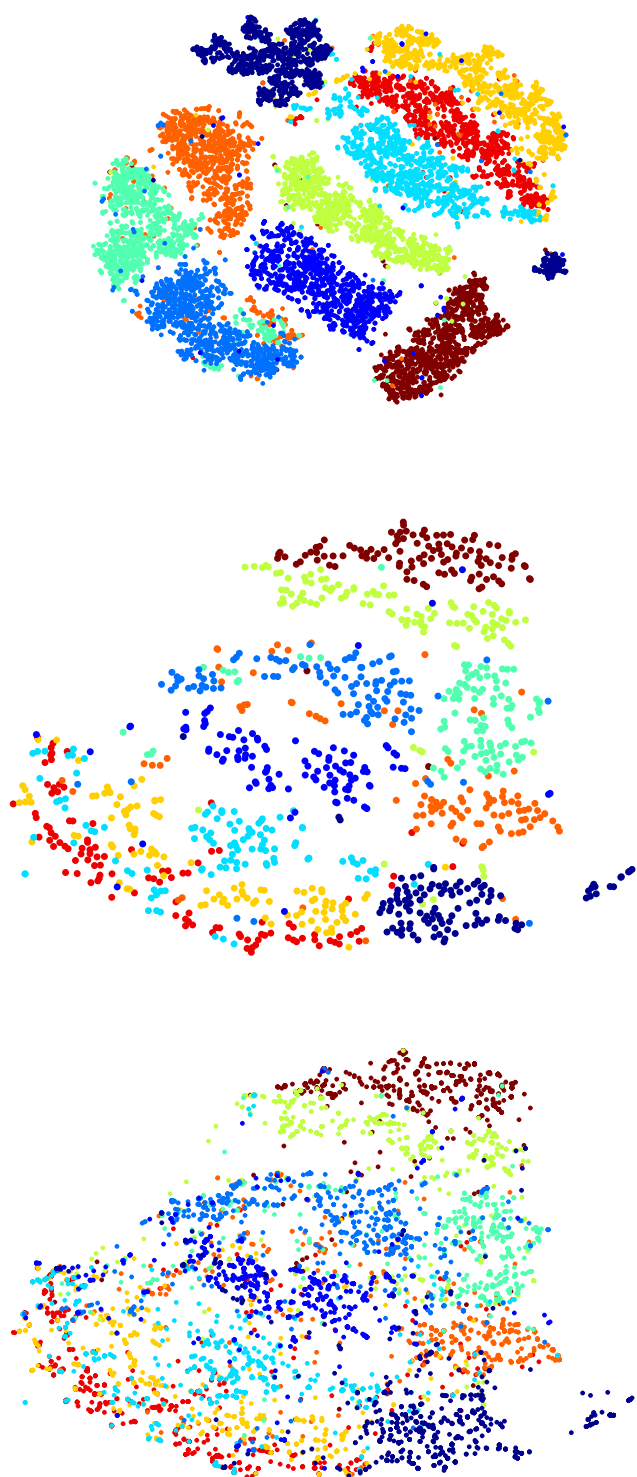


Fig. 11. Results on the usps data set. From top to bottom: projection when using t-SNE for the full data set; projection when training t-SNE on 10% of the data set; projection when extrapolating for the remaining 90% using a kernel mapping