Geodesic Generative Topographic Mapping

Raúl Cruz-Barbosa[1,2] and Alfredo Vellido[1]

[1] Universitat Politècnica de Catalunya, 08034, Barcelona, Spain
{rcruz,avellido}@lsi.upc.edu
[2] Universidad Tecnológica de la Mixteca, 69000, Huajuapan, Oaxaca, México

Abstract. Nonlinear dimensionality reduction (NLDR) methods aim to provide a faithful low-dimensional representation of multivariate data. The manifold learning family of NLDR methods, in particular, do this by defining low-dimensional manifolds embedded in the observed data space. Generative Topographic Mapping (GTM) is one such manifold learning method for multivariate data clustering and visualization. The non-linearity of the mapping it generates makes it prone to *trustworthiness* and *continuity* errors that would reduce the faithfulness of the data representation, especially for datasets of convoluted geometry. In this study, the GTM is modified to prioritize neighbourhood relationships along the generated manifold. This is accomplished through penalizing divergences between the Euclidean distances from the data points to the model prototypes and the corresponding geodesic distances along the manifold. The resulting Geodesic GTM model is shown to improve not only the *continuity* and *trustworthiness* of the representation generated by the model, but also its resilience in the presence of noise.

1 Introduction

The NLDR methods belonging to the manifold learning family model high-dimensional multivariate data under the assumption that these can be faithfully represented by a low-dimensional manifold embedded in the observed data space. This simplifying assumption may, at worst, limit the faithfulness of the generated data mapping due to either data point neighbourhood relationships that do not hold in their low-dimensional representation, hampering its *continuity*, or spurious neighbouring relationships in the representation that do not have a correspondence in the observed space, which limit the *trustworthiness* of the low-dimensional representation.

Generative Topographic Mapping (GTM) [1] is a flexible manifold learning NLDR model for simultaneous data clustering and visualization whose probabilistic nature makes possible to extend it to perform tasks such as missing data imputation [2], robust handling of outliers and unsupervised feature selection [3], or time series analysis [4], amongst others.

In the original formulation, GTM is optimized by minimization of an error that is a function of Euclidean distances, making it vulnerable to the aforementioned *continuity* and *trustworthiness* problems, especially for datasets of convoluted geometry. Such data may require plenty of folding from the GTM

H. Geffner et al. (Eds.): IBERAMIA 2008, LNAI 5290, pp. 113–122, 2008.

model, resulting in an unduly entangled embedded manifold that would hamper both the visualization of the data and the definition of clusters the model is meant to provide. Following an idea proposed in [5], the learning procedure of GTM is here modified by penalizing the divergences between the Euclidean distances from the data points to the model prototypes and the corresponding approximated geodesic distances along the manifold. By doing so, we prioritize neighbourhood relationships between points along the generated manifold, which makes the model more robust to the presence of off-manifold noise. In this paper, we first assess to what extent the resulting Geodesic GTM (or Geo-GTM) model (which incorporates the data visualization capabilities that the model proposed in [5] lacks) is capable of preserving the *trustworthiness* and *continuity* of the mapping. Then we assess whether Geo-GTM shows better behaviour in the presence of noise than its standard GTM counterpart.

2 Manifolds and Geodesic Distances

Manifold methods such as ISOMAP [6] and Curvilinear Distance Analysis [7] use the geodesic distance as a basis for generating the data manifold. This metric favours similarity along the manifold, which may help to avoid some of the distortions that the use of a standard metric such as the Euclidean distance may introduce when learning the manifold. In doing so, it can avoid the breaches of topology preservation that may occur due to excessive folding.

The otherwise computationally intractable geodesic metric can be approximated by graph distances [8], so that instead of finding the minimum arc-length between two data points lying on a manifold, we would set to find the shortest path between them, where such path is built by connecting the closest successive data points. In this paper, this is done using the K-rule, which allows connecting the K-nearest neighbours. A weighted graph is then constructed by using the data and the set of allowed connections. The data are the vertices, the allowed connections are the edges, and the edge labels are the Euclidean distances between the corresponding vertices. If the resulting graph is disconnected, some edges are added using a minimum spanning tree procedure in order to fully connect it. Finally, the distance matrix of the weighted undirected graph is obtained by repeatedly applying Dijkstra's algorithm [9], which computes the shortest path between all data samples.

3 GTM and Geo-GTM

The standard GTM is a non-linear latent variable model defined as a mapping from a low dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions generating a constrained mixture density distribution. It is defined as a generalized linear regression model:

$$\mathbf{y} = \phi(\mathbf{u})\mathbf{W}, \tag{1}$$

where ϕ are R basis functions, Gaussians in the standard formulation; $\mathbf{W}$ is a matrix of adaptive weights w_{rd}; and $\mathbf{u}$ is a point in latent space. To avoid computational intractability, a regular grid of M points $\mathbf{u}_m$ can be sampled from the latent space, which acts as visualization space. Each of them, which can be considered as the representative of a data cluster, has a fixed prior probability $p(\mathbf{u}_m) = 1/M$ and is mapped, using (1), into a low-dimensional manifold nonlinearly embedded in the data space. A probability distribution for the multivariate data $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ can then be defined, leading to the following expression for the log-likelihood:

$$L(\mathbf{W}, \beta|\mathbf{X}) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\beta}{2\pi} \right)^{D/2} \exp \left\{ -\beta/2 \|\mathbf{y}_m - \mathbf{x}_n\|^2 \right\} \right\} \tag{2}$$

where $\mathbf{y}_m$, usually known as *reference* or *prototype* vectors, are obtained for each $\mathbf{u}_m$ using (1); and β is the inverse of the noise variance, which accounts for the fact that data points might not strictly lie on the low dimensional embedded manifold generated by the GTM. The EM algorithm is an straightforward alternative to obtain the Maximum Likelihood (ML) estimates of the adaptive parameters of the model, namely $\mathbf{W}$ and β. In the E-step, the responsibilities z_{mn} (the posterior probability of cluster m membership for each data point $\mathbf{x}_n$) are computed as

$$z_{mn} = p(\mathbf{u}_m|\mathbf{x}_n, \mathbf{W}, \beta) = \frac{p(\mathbf{x}_n|\mathbf{u}_m, \mathbf{W}, \beta)p(\mathbf{u}_m)}{\sum_{m'} p(\mathbf{x}_n|\mathbf{u}_{m'}, \mathbf{W}, \beta)p(\mathbf{u}_{m'})}, \tag{3}$$

where $p(\mathbf{x}_n|\mathbf{u}_m, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{y}(\mathbf{u}_m, \mathbf{W}), \beta)$.

3.1 Geo-GTM

The Geo-GTM model is an extension of GTM that favours the similarity of points along the learned manifold, while penalizing the similarity of points that are not contiguous in the manifold, even if close in terms of the Euclidean distance. This is achieved by modifying the standard calculation of the responsibilities in (3) proportionally to the discrepancy between the geodesic (approximated by the graph) and the Euclidean distances. Such discrepancy is made operational through the definition of the exponential distribution

$$\mathcal{E}(d_g|d_e, \alpha) = \frac{1}{\alpha} \exp \left\{ -\frac{d_g(\mathbf{x}_n, \mathbf{y}_m) - d_e(\mathbf{x}_n, \mathbf{y}_m)}{\alpha} \right\}, \tag{4}$$

where $d_e(\mathbf{x}_n, \mathbf{y}_m)$ and $d_g(\mathbf{x}_n, \mathbf{y}_m)$ are, in turn, the Euclidean and graph distances between data point $\mathbf{x}_n$ and the GTM prototype $\mathbf{y}_m$. Responsibilities are redefined as:

$$z_{mn}^{geo} = p(\mathbf{u}_m|\mathbf{x}_n, \mathbf{W}, \beta) = \frac{p'(\mathbf{x}_n|\mathbf{u}_m, \mathbf{W}, \beta)p(\mathbf{u}_m)}{\sum_{m'} p'(\mathbf{x}_n|\mathbf{u}_{m'}, \mathbf{W}, \beta)p(\mathbf{u}_{m'})}, \tag{5}$$

where $p'(\mathbf{x}_n|\mathbf{u}_m, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{y}(\mathbf{u}_m, \mathbf{W}), \beta)\mathcal{E}(d_g(\mathbf{x}_n, \mathbf{y}_m)^2|d_e(\mathbf{x}_n, \mathbf{y}_m)^2, 1)$. When there is no agreement between the graph approximation of the geodesic distance and the Euclidean distance, the value of the numerator of the fraction within the exponential in (4) increases, pushing the exponential and, as a result, the modified responsibility, towards smaller values, i.e., punishing the discrepancy between metrics. Once the responsibility is calculated in the modified E-step, the rest of the model's parameters are estimated following the standard EM procedure.

3.2 Data Visualization Using Geo-GTM

The GTM was originally defined as a probabilistic alternative to Self-Organizing Maps (SOM). As a result, the data visualization capabilities of the latter are fully preserved and even augmented by the former. The main advantage of GTM and any of its extensions over general finite mixture models consists precisely on the fact that both data and results can be intuitively visualized on a low dimensional representation space.

Each of the cluster representatives $\mathbf{u}_m$ in the latent visualization space is mapped, following Eq. (1), into a point $\mathbf{y}_m$ (the center of a mixture component) belonging to a manifold embedded in data space. It is this mapping (and the possibility to invert it) what provides Geo-GTM with the data visualization capabilities that the alternative Manifold Finite Gaussian Mixtures model proposed in [5] lacks. Given that the posterior probability of every Geo-GTM cluster representative for being the generator of each data point, or responsibility z_{mn}^{geo}, is calculated as part of the modified EM algorithm, both data points and cluster prototypes can be visualized as a function of the latent point locations as the mean of the estimated posterior distribution:

$$\mathbf{u}_n^{mean} = \sum_{m=1}^{M} \mathbf{u}_m z_{mn}^{geo}, \tag{6}$$

4 Experiments

Geo-GTM was implemented in MATLAB®. For the experiments reported next, the adaptive matrix $\mathbf{W}$ was initialized, following a procedure described in [1], as to minimize the difference between the prototype vectors $\mathbf{y}_m$ and the vectors that would be generated in data space by a partial Principal Component Analysis (PCA). The inverse variance β was initialised to be the inverse of the 3^{rd} PCA eigenvalue. This initialization ensures the replicability of the results. The latent grid was fixed to a square layout of approximately $(N/2)^{1/2} \times (N/2)^{1/2}$, where N is the number of points in the dataset. The corresponding grid of basis functions was equally fixed to a 5×5 square layout for all datasets.

The goal of the experiments is threefold. Firstly, we aim to assess whether the proposed Geo-GTM model could capture and visually represent the underlying structure of datasets of smooth but convoluted geometry better than the

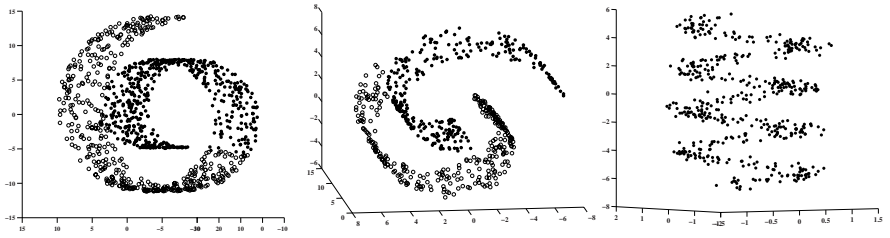

Fig. 1. The three datasets used in the experiments. (Left): *Swiss-Roll*, where two contiguous fragments are identified with different symbols in order to check manifold contiguity preservation in Fig. 3. (Center): *Two-Spirals*, again with different symbols for each of the spiral fragments. (Right): *Helix*.

standard GTM. Secondly, we aim to quantify the faithfulness of the generated mappings. Finally, we aim to evaluate the capability of Geo-GTM to uncover the underlying structure of the data in the presence of noise, and compare its performance with that of the standard GTM.

4.1 Results and Discussion

Three artificial 3-dimensional datasets, represented in Fig. 1, were used in the experiments that follow. The first one is *Swiss-Roll*, consisting on 1000 randomly sampled data points generated by the function: $(x_1, x_2) = (t\cos(t),\ t\sin(t))$, where t follows a uniform distribution $\mathcal{U}(3\pi/2, 9\pi/2)$ and the third dimension follows a uniform distribution $\mathcal{U}(0, 21)$. The second dataset, herein called *Two-Spirals*, consists of two groups of 300 data points each that are similar to *Swiss-Roll* although, this time, the first group follows the uniform distribution $\mathcal{U}(3\pi/4, 9\pi/4)$, while the second group was obtained by rotating the first one by 180 degrees in the plane defined by the first two axes and translating it by 2 units along the third axis. The third dataset, herein called *Helix*, consists of 500 data points that are images of the function $\mathbf{x} = (\sin(4\pi t), \cos(4\pi t), 6t - 0.5)$, where t follows $\mathcal{U}(-1, 1)$. These data are contaminated with a small level of noise. Also, and specifically for the experiments to assess the way the models deal with the presence of noise, Gaussian noise of zero mean and increasing standard deviation, from $\sigma = 0.1$ to $\sigma = 0.5$, was added to a noise-free version of *Helix* to produce the 5 datasets represented in Fig. 2.

The posterior mean distribution visualization maps for all datasets are displayed in Figs. 3 to 5. Geo-GTM, in Fig. 3, is shown to capture the spiral structure of *Swiss-Roll* far better than standard GTM, which misses it at large and generates a poor data visualization with large overlapping between non-contiguous areas of the data.

A similar situation is reflected in Fig. 4: The two segments of *Two-Spirals* are neatly separated by Geo-GTM, whereas the standard GTM suffers a lack of contiguity of the segment represented by circles as well as overlapping of part of the data of both segments.

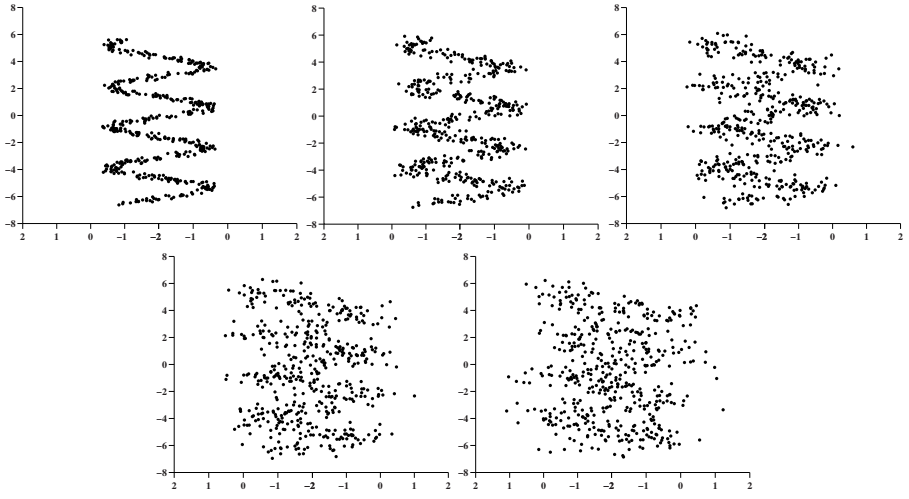

Fig. 2. The five noisy variations of *Helix* used in the experiments. From left to right and top to bottom, with increasing noise of standard deviation from $\sigma = 0.1$ to $\sigma = 0.5$.

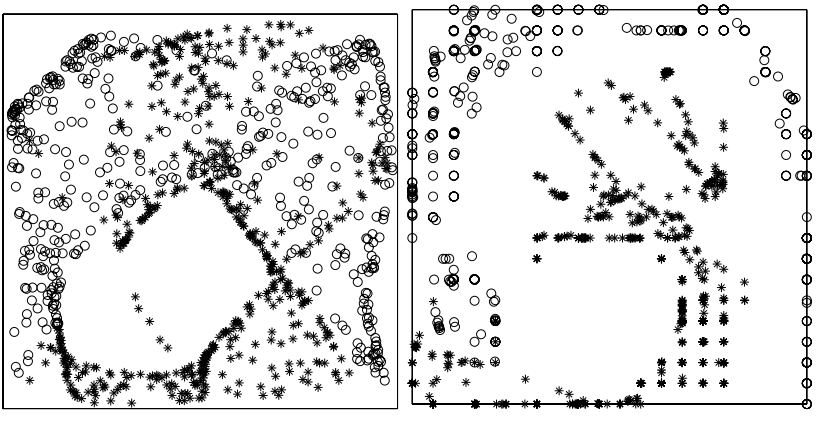

Fig. 3. Data visualization maps for the *Swiss-Roll* set. (Left): standard GTM; (right): Geo-GTM.

The results are even more striking for *Helix*, as shown in Fig. 5: the helicoidal structure is neatly revealed by Geo-GTM, whereas it is mostly missed by the standard GTM. The former also faithfully preserves data continuity, in comparison to the breaches of continuity that hinder the visualization generated by the latter.

In order to evaluate and compare the mappings generated by GTM and Geo-GTM, we use the *trustworthiness* and *continuity* measures developed in [10]. *Trustworthiness* is defined as:

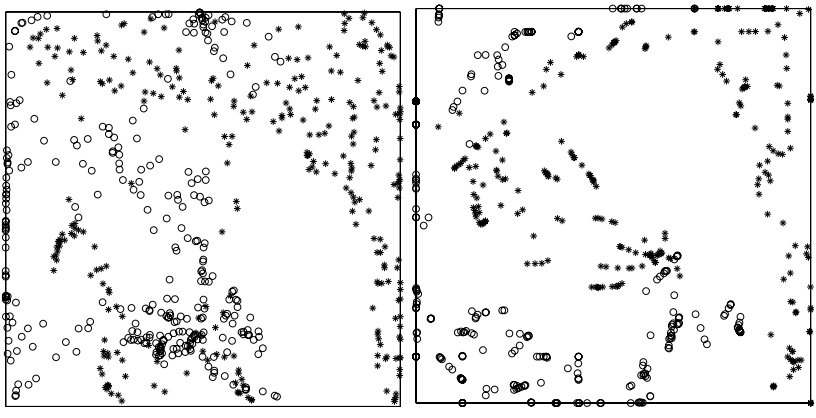

Fig. 4. Visualization maps for the *Two-Spirals* set. (Left): standard GTM; (right): Geo-GTM.

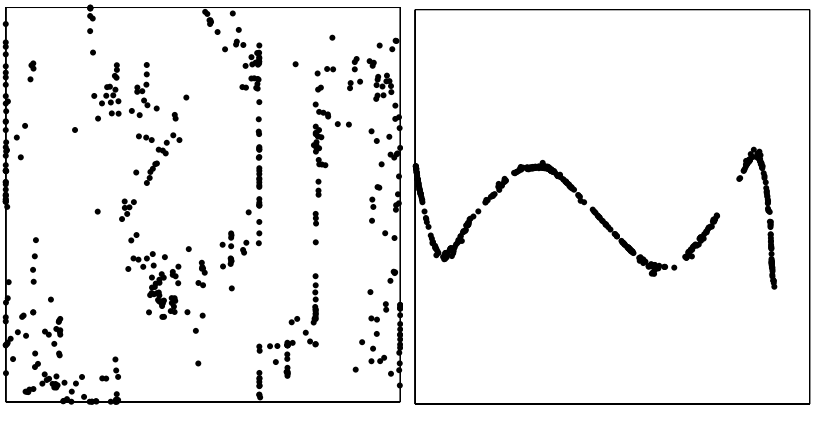

Fig. 5. Data visualization maps for the *Helix* set. (Left): standard GTM; (right): Geo-GTM.

$$T(K) = 1 - \frac{2}{NK(2N - 3K - 1)} \sum_{i=1}^{N} \sum_{x_j \in U_K(x_i)} (r(x_i, x_j) - K), \tag{7}$$

where $U_k(x_i)$ is the set of data points x_j for which $x_j \in \hat{C}_K(x_i) \wedge x_j \notin C_K(x_i)$ and $C_K(x_i)$ and $\hat{C}_K(x_i)$ are the sets of K data points that are closest to x_i in the observed data space and in the low-dimensional representation space, respectively. *Continuity* is in turn defined as:

$$Cont(K) = 1 - \frac{2}{NK(2N - 3K - 1)} \sum_{i=1}^{N} \sum_{x_j \in V_K(x_i)} (\hat{r}(x_i, x_j) - K), \tag{8}$$

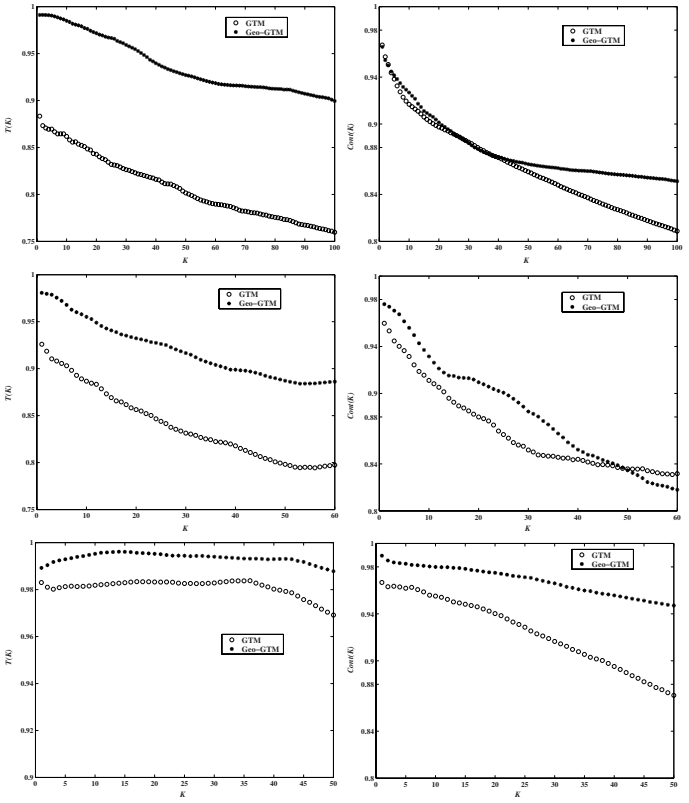

Fig. 6. Trustworthiness (left column) and continuity (right column) for (top row): *Swiss-Roll*, (middle row): *Two-Spirals*, and (bottom row): *Helix*, as a function of the neighbourhood size K

where $V_K(x_i)$ is the set of data points x_j for which $x_j \notin \hat{C}_K(x_i) \wedge x_j \in C_K(x_i)$. The terms $r(x_i, x_j)$ and $\hat{r}(x_i, x_j)$ are the ranks of x_j when data points are ordered according to their distance from the data vector x_i in the observed data space and in the low-dimensional representation space, respectively, for $i \neq j$.

The measurements of *trustworthiness* and *continuity* for all datasets are shown in Fig. 6. As expected from the visualization maps in Figs. 3-5, the Geo-GTM mappings are far more trustworthy than those generated by GTM for neighbourhoods of any size across the analyzed range. The differences in continuity preservation are smaller although, overall, Geo-GTM performs better than GTM model, specially with the noisier *Helix* dataset.

We finally evaluate, through some preliminary and illustrative experiments, the capability of Geo-GTM to uncover the underlying structure of the data in the presence of noise, comparing it with that of the standard GTM. We quantify it using the log-likelihood (2), as applied to a test dataset consisting of 500 randomly sampled data points from a noise-free version of *Helix*. For further

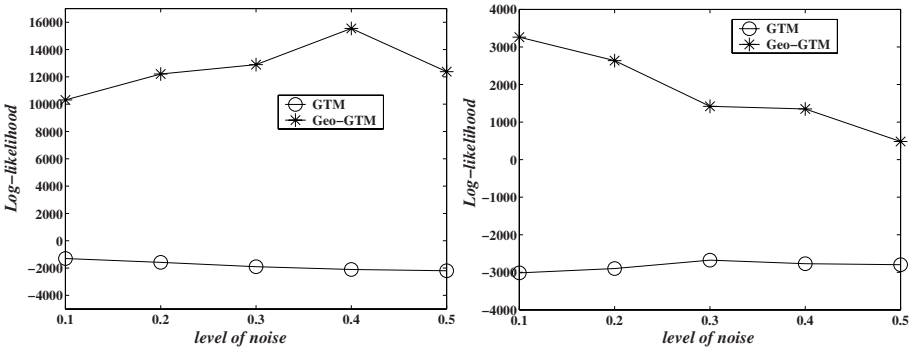

Fig. 7. Test log-likelihood results for the *Helix* (left) and *Two-Helix* (right) datasets, for increasing levels of added uninformative noise

testing, we repeat the experiment with noisy variations of a basic dataset, herein called *Two-Helix* consisting of two sub-groups of 300 data points each, which are, in turn, images of the functions $\mathbf{x}1 = (sin(4\pi t), cos(4\pi t), 6t - 0.5)$ and $\mathbf{x}2 = (-sin(4\pi t), -cos(4\pi t), 6t - 0.5)$, where t follows $\mathcal{U}(-1, 1)$. This is, in fact, a DNA-like shaped duplication of the *Helix* dataset. The corresponding results are shown in Fig. 7.

Remarkably, Geo-GTM is much less affected by noise than the standard GTM, as it recovers with much higher likelihood the underlying noise-free functions. This corroborates the visualization results reported in Fig. 5, in which the standard GTM generates a far less faithful representation of the underlying form and with breaches of continuity. This is probably due to the fact that Geo-GTM favours directions along the manifold, minimizing the impact of off-manifold noise.

5 Conclusion

A variation of the NLDR manifold learning GTM model, namely Geo-GTM, has been shown in this study to be able to faithfully recover and visually represent the underlying structure of datasets of smooth but convolute geometries. It does so by limiting the effect of manifold folding through the penalization of the discrepancies between inter-point Euclidean distances and the approximation of geodesic distances along the model manifold. As a byproduct of this approach, Geo-GTM avoids, at least partially, the problem of overfitting by penalizing off-manifold neighbouring relationships between data points. The reported experiments also show that Geo-GTM has recovered the true underlying data structure far better than the standard GTM, even in the presence of a considerable amount of noise. Future research should extend these experiments to a wider selection of datasets, both synthetic and real. It should also investigate, in wider detail, alternative approaches to graph generation as an approximation to geodesic distances. For this, we resorted in this paper to the K-rule, but other

approaches could be considered, such as the ϵ-rule, the τ-rule, or even the more sophisticated Data- and Histogram- rules [11].

Acknowledgements. Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish MICINN and acknowledges funding from the CICyT research project TIN2006-08114. Raúl Cruz-Barbosa acknowledges SEP-SESIC (PROMEP program) of México for his PhD grant.

References

1. Bishop, C.M., Svensén, M., Williams, C.K.I.: The Generative Topographic Mapping. Neural Computation 10(1), 215–234 (1998)
2. Vellido, A.: Missing data imputation through GTM as a mixture of t-distributions. Neural Networks 19(10), 1624–1635 (2006)
3. Vellido, A., Lisboa, P.J.G., Vicente, D.: Robust analysis of MRS brain tumour data using t-GTM. Neurocomputing 69(7-9), 754–768 (2006)
4. Olier, I., Vellido, A.: Advances in clustering and visualization of time series using GTM Through Time. Neural Networks (accepted for publication)
5. Archambeau, C., Verleysen, M.: Manifold constrained finite Gaussian mixtures. In: Cabestany, J., Gonzalez Prieto, A., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 820–828. Springer, Heidelberg (2005)
6. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
7. Lee, J.A., Lendasse, A., Verleysen, M.: Curvilinear Distance Analysis versus Isomap. In: Proceedings of European Symposium on Artificial Neural Networks (ESANN), pp. 185–192 (2002)
8. Bernstein, M., de Silva, V., Langford, J., Tenenbaum, J.: Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, CA (2000)
9. Dijkstra, E.W.: A note on two problems in connection with graphs. Numerische Mathematik 1, 269–271 (1959)
10. Venna, J., Kaski, S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 485–491. Springer, Heidelberg (2001)
11. Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, New York (2007)