Probabilistic Elastic Embedding Model: Comparison of Alternative Models

Rodolphe Priam

Probabilistic Elastic Embedding Model: Comparison of Alternative Models*

R. Priam†

June 13, 2018 (updated 17/09/2021)

Abstract

In data visualization, Elastic Embedding adds an exponential penalty to an Euclidean criterion. It is able to separate the natural classes but its lacks a probabilistic generative setting which brings more flexibility to the modeling and the inference. Hence, it is proposed a new generative interpretation of Elastic Embedding which is closely related to LargeVis. Numerical experiments compare the proposed model and several alternative ones via two new visual indicators among different approaches.

1 Introduction

In data visualization, the observations in the available sample are vectorial: the rows or the columns of numerical data matrices or tables. A family of methods is dedicated to the symmetric numerical matrices which contain the distances or similarities between high-dimensional data vectors. An extensive literature exists and diverse approaches have been developped until today in this domain of research but they can appear in very different domains of research even if they are closely related: visualization, word embedding, statistical analysis and clustering analysis. Currently, the method *t-Distributed Stochastic Neighbor Embedding* (t-SNE or tSNE) [1] is the state of art for large datasets, it is non generative and scaled well for large datasets. For a generative/probabilistic basis, it can be cited for instance *Probabilistic Principal Component Analysis* [2], *Glove* [3], and *LargeVis* [4] with a very good scaling. Nextafter, this section presents the notation, the purpose and the plan of the paper.

Data notation and reduction: Let have the available high-dimensional data as a set of data vectors in a space with M dimensions as follows:

$$\mathscr{X} = \{\mathbf{x}_i \in R^M; 1 \leq i \leq N\}.$$

Let define $W = (w_{ij})$ from the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, for instance $d_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|$, with $w_{ij} = d_{ij}$, or more generally a function of d_{ij}. Typically, a weighted nearest neighboors graph such as a heat matrix with for instance $w_{ij} = e^{-0.5d_{ij}^2/\tau}$ for $\tau > 0$ when d_{ij} is enough small, and $w_{ij} = 0$ otherwise. A weighted graph $G = (V, E, W)$ is defined from V, E, and W which stands respectively for the set of vertices, edges and weights: an edge e_{ij} comes from from a pair of vertices (v_i, v_j) in the graph of nearest neighboors. It is also denoted $\bar{E}$ for the set of pairs of vertices that are not neighboors.

The purpose of a reduction is to summarize $\mathscr{X}$ by finding relevant lower dimensional representations:

$$\mathscr{Y} = \{\mathbf{y}_i \in R^S; 1 \leq i \leq N\}.$$

Generally $S = 2$ as the visualization appears in the two dimensional plane, even if three dimensions or even more remain possible. Existing models consider an embedding of low dimensional positions via their pairwise distances/similarities plus bias/intercept terms. The parameterization approximate the true distances d_{ij} between data pairs $(\mathbf{x}_i, \mathbf{x}_j)$ via the distances between pairs of S-dimensional lower position vectors $(\mathbf{y}_i, \mathbf{y}_j)$ denoted δ_{ij}. Let also

*A nearly similar version of this document was sent to review previously from 2018, and after a shorter version without section 3.3 was accepted at a conference in 2020 but canceled.

†rpriam@gmail.com

denote the Poisson mass functions as follows, $\varphi_{\text{PSN}}(.,\delta) = \frac{1}{.!}(.)^{e^\delta} e^{-e^\delta}$.

Purpose of the paper: Among the existing non parametric methods, the Elastic embedding [5] has been shown to perform well on several datasets such that it may be extended for even better results. EE is defined via a penalized Eudlidean criterion which aims at retrieving the true pairwise distances between each couple of data through new variables which represent the data in a lower dimensional space. Parametric probabilistic models for visualization are generally more flexible than ad hoc criterion and the probabilities have an intuitive interpretation, hence introducing such framework for EE is of main interest.

Plan: Section 2 presents EE, its probabilistic version, links with the current literature and its illustration with real data. Sections 3 presents an analytical evaluation of the model under some hypothesis with a new framework for the comparisons. Section 4 concludes the paper with future perspectives.

2 Generative Elastic Embedding

In this section, Elastic Embedding is reviewed before introducing a new parameterization in order to deduce a probabilistic model for visualization.

2.1 Elastic Embedding

The model EE was introduced as an alternative of SNE [6] for dimensionality reduction. The author Miguel A. Carreira-Perpinan notices a link between SNE [6] and Laplacian eigenmaps (LE) [7]. The SNE criterion is rewritten into two parts: a) the first term is exactly equal to the LE objective except that normalised affinities would be considered, and b) the second term encourages latent points to be far away with a log of a sum of exponential quantities. The author explains that the SNE criterion induces that the data vectors in a same neighboor are mapped into a same area of the low dimensional view and also separates every projections points instead of focusing on the separation of the clusters themself.

In the more formal setting, when w_{ij}^+ and w_{ij}^- stands for given weights, and $\delta_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|^2$, the objective function of EE is as follows,

$$C_{EE}(\mathscr{Y}, \gamma) = \sum_{i,j} w_{ij}^+ \delta_{ij} + \gamma \sum_{i,j} w_{ij}^- \exp(-\delta_{ij}) .$$

For choosing the input values for the two sets of weights w_{ij}^+ and w_{ij}^-, many approaches are possible as explained by the author of the method.

- In the original paper of EE it is choosen the following one:

$$\begin{aligned} w_{ij}^+ &= e^{-0.5 d_{ij}^2/\tau} \\ w_{ij}^- &= \begin{cases} d_{ij}^2 & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases} \end{aligned} .$$

  It is underlined in the research paper on EE that other weighting are possible such as sparser graphs.

- For sparser weights, it can be noticed that for $\mathbf{x}_i$ and $\mathbf{x}_j$ enough near the weight w_{ij}^- vanishes, hence summing for only non neighboors pairs for the original space leads to a very similar criterion. In the same idea for $\mathbf{x}_i$ and $\mathbf{x}_j$ enough far away the weight w_{ij}^+ vanishes hence summing for only neighboors pairs leads also to a similar criterion. These approximations have for corresponding criterion $\tilde{C}_{EE}$ with new weights $\tilde{w}_{ij}^+$ and $\tilde{w}_{ij}^-$ as follows,

$$\begin{aligned} \tilde{w}_{ij}^+ &= \begin{cases} w_{ij}^+ & \text{if } j \in \mathscr{N}_i \\ 0 & \text{otherwise} \end{cases} \\ \tilde{w}_{ij}^- &= \begin{cases} w_{ij}^- & \text{if } j \notin \mathscr{N}_i \\ 0 & \text{otherwise} \end{cases} \end{aligned} .$$

  This difference of modeling for each part of the graph E and $\bar{E}$ recalls recents models in the literature, in particular LargeVis. A limit of the approximation is to not separate well projection in a given cluster but for very large dataset, this problem may be limited because viewing separately each point appears not relevant for a human and only zooming at a given area of the map might suffer from this approximation. Some authors add an additional pernalty for the local projection but this is not studied herein and only this criterion is considered hereafter, with same name EE.

2.2 Generative version

For a generative setting of EE, it is proposed to rewrite the approximated criterion denoted $\tilde{C}_{EE}$ with the constant[1] C_w. The weights are supposed w_{ij}^+ integer and w_{ij}^- binary, such that for $\alpha > 0$ small,

$$\begin{aligned} & -\tilde{C}_{EE}(\mathscr{Y},\gamma) - C_w \\ = & + \sum_{(i,j)\in E} w_{ij}^+ \log e^{-\delta_{ij}+\log\alpha} - \gamma \sum_{(i,j)\in \bar{E}} w_{ij}^- e^{-\delta_{ij}} - C_w \\ \approx & + \sum_{(i,j)\in E} \log\{e^{-\delta_{ij}+\log\alpha}\}^{w_{ij}^+} - \alpha w_{ij}^+ e^{-\delta_{ij}} - \log w_{ij^+}! \\ & + \gamma \sum_{(i,j)\in \bar{E}} \log\{e^{-e^{-\delta_{ij}}}\} \\ = & \log\left[\left\{ \prod_{(i,j)\in E} \frac{(\tilde{\delta}_{ij})^{w_{ij}} e^{-\tilde{\delta}_{ij}}}{w_{ij}!} \right\} \left\{ \prod_{(i,j)\in \bar{E}} e^{-\tilde{\delta}_{ij}} \right\}^{\gamma} \right]. \end{aligned}$$

Hence, it is obtained,

$$\operatorname{argmin}_{\mathscr{Y}} \tilde{C}_{EE}(\mathscr{Y},\gamma) \approx \operatorname{argmax}_{\mathscr{Y}} \tilde{L}_{EE}(\mathscr{Y},\gamma),$$

when,

$$\tilde{\delta}_{ij} = \begin{cases} e^{-\delta_{ij}+\log w_{ij}^-} & \text{denoted } \tilde{\delta}_{ij}^- \text{ for } \bar{E} \\ e^{-\delta_{ij}+\log\alpha} & \text{denoted } \tilde{\delta}_{ij}^+ \text{ for } E \end{cases}$$

More formally, this is the behavior of the Poisson mass distribution associated to Euclidean distance which is in stake here,

Let denote δ a real and w a positive integer. Then the logarithm of the mass distribution $\varphi_{PSN}(w, e^{-\delta+\log\alpha})$ behaves like the limit value $-w\delta$ for a real enough small α.

This is a consequence of the logarithm of φ_{PSN} with the introduced parameterization is equal to $-w\delta + w\log\alpha - \alpha e^{-\delta} - \log w!$. This induces that only the quantity $-w\delta$ is in stake in the maximization and is minimized as wanted for EE. The proposed likelihood for EE is then,

$$\begin{aligned} & \tilde{L}_{EE}^{\alpha}(\mathscr{Y},\gamma) \\ = & \prod_{(i,j)\in E} \varphi_{\text{PSN}}(w_{ij}^+, \tilde{\delta}_{ij}^+) \prod_{(i,j)\in \bar{E}} \varphi_{\text{PSN}}(0, \gamma\tilde{\delta}_{ij}^-). \end{aligned}$$

Note that this is enough general to apply to some other methods for visualization in order to define a generative version. This model can be seen as a mixture between the observed nodes and the non observed nodes with a classifying likelihood. It is named genEE and the inference for the parameters $\mathscr{Y}$ is discussed next subsection.

[1] $C_w = -\sum_{(i,j)\in E} \log w_{ij}^+ - \log\alpha \sum_{(i,j)\in E} w_{ij}^+$

2.3 Links with the literature

Several models are closely related to genEE as latent position models have a long history in the statistical and computational literature. The generative model for correspondence analysis [8] is based on a Poisson with a parameterization $e^{\mathbf{y}_i^T \mathbf{y}_j + b_i + b_j}$. The scalar quantities aggregated in $\mathscr{B} = (b_i)_i$ are fixed and equal to the logarithm of the margin sums normalized to one: the similarity is different to the Euclidean distance and less flexible than in Glove [9]. Here $\tilde{\delta}_{ij}$ is written with a scalar product instead of the less flexible Euclidean distance. This model is also defined for a rectangular matrix of positive integer and was shown to lead to almost perfectly equal estimation of the latent position than the matricial correspondence analysis [10]. In genEE, a criterion very similar to LargeVis (LV) is recognized except that a Poisson is the foundation for the underlying distributional hypothesis instead of a Bernouilli distribution, as suggested in [11]. The probabilistic interpretation with w_{ij} an integer remains the product between the likelihood of the observed edges with the likelihood of the non observed edges with a weighting γ for regulating the importance of each distribution. As a remark, alternative distributions such as multinomial, negative poisson, binomial or even normal may be also possible. The main difference with genEE is that the weights are included in the model hence a fully probabilistic approach may be possible if the fitting is statistically relevant. The model genEE may be also able to exactly behave like LV when a suitable parameterization is chosen for the expression of the expectations.

2.4 Empirical illustration

In this section, the models are compared with a real dataset in order to discuss the quality of the clusters obtained on the nonlinear embedding by several methods. Several indicators are considered for the evaluation of the compactness and the distance between clusters.

- *Data:* The MNIST dataset is a subset of a larger set available from NIST, it counts 60000 examples in the training set and 10000 examples in the test set. The digits have been size-normalized and centered in a fixed-size image. The original bicolor images from NIST were first coded in

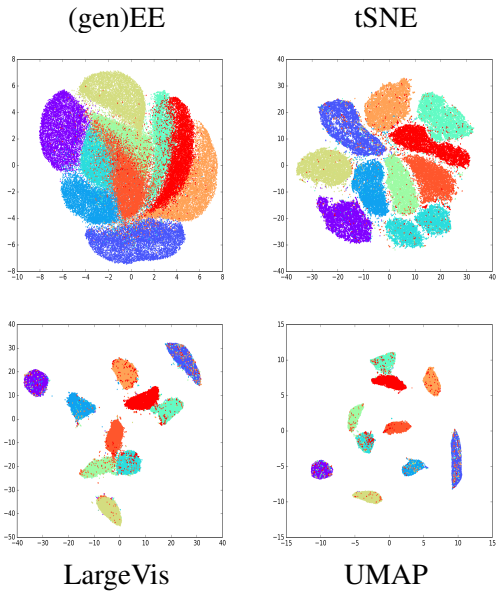

Figure 1: Nonlinear embeddings for MNIST.

20x20 arrays with grey level and then centered in 28x28 arrays.

- *Experimental settings:* For the learning of the quantities $\mathscr{Y}$ in EE, a sequential algorithm is processed, this approach is recently widely studied. In particular, it is implemented the probabilistic edge sampling [4] and negative sampling [9] for an approximate stochastic gradient descent algorithm without normalization. For tSNE, LargeVis and UMAP [12], the available packages in language python have been used.

- *Empirical results:* On the Figure 2, the 10 classes are almost perfectly separated on the map from the recents methods tSNE, LargeVis and UMAP with large frontiers for the two last ones. EE performs less well for this dataset with smaller frontiers around the clusters.

Next section is dedicated to explaining the differences between the results of the different methods.

3 Numerical comparison of the criteria

A study of the behavior of the criteria under simple hypothesis are presented in this subsection in order to help the understanding of their empirical difference in practice. A trade-off between the clusters compactness, clusters separation plus the quality of the global and local projection needs to be found in order to bring the best map. According to the indicators in this section, two properties of the criteria a checked. For the compactness of a cluster, the coefficient of variation (CV) has to be enough small. For the separation between the clusters, either the clusters are degenerated into their center either their compactness has an influence with the frontiers.

3.1 Distance between two clusters

Following [13], the distance between two subgraphes can be modeled via the approximation that every nodes in each subgraph have same final positions. Let denote the number of nodes in first subgraph n_1 and n_2 the number for the second subgraph, and the numbe of edges n_e, with $n_e \ll n_1 n_2$ and $f = n_1 n_2 / n_e$. The criterion has the following form:

$$u(\delta) = n_e\, g(\delta) + n_1 n_2\, h(\delta)\,.$$

The function $g(.)$ and $h(.)$ depends on the criterion in stake, several functions are presented in [14] for older methods. Note that the function $g(.)$ and $h(.)$ have the following proporties: g is increasing and h is decreasing from a value $\delta > 0$ hence $h'(\delta) > 0$ and $h'(\delta) < 0$ while the second order derivatives are supposed positive for insuring a solution. The functions $g(.)$ and $h(.)$ lead to competitive results when they lead to small clusters and large frontiers between the clusters.

For the distances, this framework results into the solutions a) for [15] such as $\hat{\delta}^2 \log \hat{x} = f$, b) $\hat{\delta} = \sqrt[3]{f}$ in [16], c) $\hat{\delta} = \sqrt[4]{f}$ in [17] and $\hat{\delta} = f$ for LinLog in [13, 14]. This leads to justify the better results for the method LinLog with an empirical illustration with perfectly symmetric clusters where the frontiers was shown to be more apparent. For Elastic Embedding with full weigths, it may be found $\hat{\delta} = \log(\gamma f)$. A limit of this criterion is when only the nearest neighboors are involved for the local projection, then the function $g(.)$ may not anymore be involved too. A complement of this study is the influence of the components of the criteria to the compactness of the clusters which is underlined hereafter.

3.2 Compactness of the clusters: indicator

For a better idea of how behave a criterion from the point of view of the visualization of clusters it may be observed the first moments of the part related to the local projection and the global projection. To our knowledge, this question has not been studied in the literature for the quality of a visualization in a general situation.

- *Hypothesis:* it is supposed the Gaussianity of the clusters with ellipsoidal shapes. The random variables in stake are,

$$\delta_{ij} = \| \mathbf{y}_i - \mathbf{y}_j \|^2$$

With $\mathscr{G}(\mu,\Sigma)$ is for a Gaussian random variable with mean μ and variance Σ, the distribution is as follows,

$$\mathbf{y}_i \sim \mathscr{G}(\mu, \tfrac{1}{2}\Sigma).$$

When $\mathbf{y}_i$ and $\mathbf{y}_j$ are independent, $\mathbf{y}_i - \mathbf{y}_j \sim \mathscr{G}(0,\Sigma)$ hence their difference has a similar distribution but with a zero expectation and a double variance. For diagonal matrices $\Sigma = \text{diag}_k(\sigma_k^2)$ with non null elements σ_k^2 for the variances per dimension, according to [18], the distribution of δ_{ij} is the sum of S random variables, say informally,

$$\delta_{ij} \sim \textstyle\sum_k \Gamma(0.5, 2\sigma_k^2).$$

In the simple case when $\sigma_k = 1$, it is known that the Euclidean distance follows a χ^2 hence the variance is $2S$ and the mean is S. For the more general case, under the hypothesis of diagonal Gaussianity for $\mathbf{y}_i$, the expectation and variance of δ_{ij} are respectively,

$$\begin{array}{rcccl} \mu_\delta & = & \mathbb{E}[\delta_{ij}] & = & \sum_k \sigma_k^2 \\ \sigma_\delta^2 & = & \mathbb{V}ar[\delta_{ij}] & = & 2\sum_k \sigma_k^4. \end{array}$$

It is retrieved the case of the χ^2 when it is replaced σ_k by 1 with corresponding mean and variance.

- *Coefficient of variation (CV):* this is a measure of the dispersion of a probability distribution. It is written as the ratio of the standard deviation to the mean. A function with smaller coefficient of variation may lead to more compact clusters because the values of the function has less variations during the projection where the positions $\mathbf{y}_i$ are constructed. The expression of the indicator for the measure of the scattering is as follows,

$$\begin{array}{rcl} I_{CV} & = & \frac{\sqrt{\mathbb{V}ar[g(\delta_{ij})]}}{\mathbb{E}[g(\delta_{ij})]} \\ & \approx & \frac{\sqrt{(g'(\mu_\delta))^2\sigma_\delta^2 + \frac{(g''(\mu_\delta))^2\sigma_\delta^4}{4}}}{g(\mu_\delta) + \frac{g''(\mu_\delta)\sigma_\delta^2}{2}}. \end{array}$$

For a given function $g(\delta_{ij})$, a functional expansion suggests an approximate value. This indicator is for instance estimated via simulation of Gaussian distributions for $\mathbf{y}_i$, and the computations of the resulting distance. It is computed via simulation of (1000 here) vectors for $S = 2$ for some Gaussian distribution, in the Table 1. Note that the function $e^{atan(\sqrt{\delta_{ij}})}$ is also evaluated because of its derivative proportional to $(1+\delta_{ij})^{-1}$ which is very similar to the logarithmic case for $\log(1+\delta_{ij})$. The table 1 presents the indicators for different shapes of projection: it may confirm the compacity of the clusters for LargeVis in comparison to other methods with a smaller indicator. The alternative functions with an even smaller indicator I_{CV} are expected candidates to behave well at least for small datasets and well separated classes. Next a more general model is proposed for evaluating a criterion when the clusters have spherical shapes.

Table 1: Indicators of compactness with $\Sigma = diag(a,b)$.

Function g or h	I_{CV} $a=1.0$ $b=1.0$	I_{CV} $a=0.5$ $b=1.0$	I_{CV} $a=2.0$ $b=1.0$
δ_{ij}	1.00	1.04	1.05
$\log(1+e^{-\delta_{ij}})$	1.24	1.04	1.50
$\log(1+\delta_{ij})$	0.55	0.58	0.51
$e^{atan(\sqrt{\delta_{ij}})}$	0.24	0.24	0.23
$e^{-1/(0.1+\delta_{ij})}$	0.44	0.50	0.38
$\log(\delta_{ij}) - \log(1+\delta_{ij})$	-1.22	-1.11	-1.36
$e^{-\delta_{ij}}$	1.34	1.13	1.62
$-\log(\delta_{ij})$	-1.57	-2.47	-1.11

3.3 Compactness of the clusters: model

The generalized criterion of EE is approximated for binary weigths, with $E = \cup_\ell E_\ell$, as follows,

$$\begin{aligned} & C_{EE}(\mathscr{Y}, \gamma) \\ = & \sum_{(i,j)\in E} g(\delta_{ij}) + \gamma \sum_{(i,j)\in \bar{E}} h(\delta_{ij}) \\ \approx & \sum_k \sum_{(i,j)\in E_\ell} g(\delta_{ij}) + \gamma \sum_{(i,j)\in \bar{E}} h(\delta_{ij}) \\ \approx & \sum_\ell n_\ell \int g(\delta) f_\ell^+(\delta) d\delta + \gamma n_e \int h(\delta) f^-(\delta) d\delta \,. \end{aligned}$$

Here it is clear that under an hypothesis of clustering, the distances δ_{ij} are considered separately in each cluster for the positive interactions (left part) while an unique cluster is considered for the negative interactions (right part). As seen in the paragraph above, under Gaussianity, the densities f_ℓ^+ and f^- are the Gamma depending on the variances of δ_{ij}, respectively in E_ℓ and $\bar{E}$.

The hypothesis of indendence between the distances may be strong but is taken nextafter. For comparison purpose, it is supposed only two clusters and all the variances per dimension equal to σ^2 hence in each cluster the distribution of δ_{ij} is chosen:

$$f_\ell^+ \sim \Gamma(1, 2\sigma^2) \,.$$

For modeling the distribution outside the clusters it is taken an ad hoc value equal to $2.1\,\sigma^2$ which is the double of one cluster plus some variability from the frontier such that the distribution involved is chosen:

$$f^- \sim \Gamma(1, 4.2\sigma^2) \,.$$

Even if this choice is not optimal because the variance for the interactions may vary from a model to another, it leads to a first informative result. The optimization for a projection is reduced at finding the value of σ^2 which makes minimal the simplified criterion for several choice of function $g(.)$ and $h(.)$.

The first results with such model is very encouraging. The obtained curves of the criteria against the variable σ^2 are shown just after for Elastic Embedding (EE), LinLog (LL) and LargeVis (LV), when $n_1 = n_2 = 50$ and $n_2 = 350$ while $\gamma = 1$.

The proposed modeling of a criterion for visualization is able to show without any simulation of the related method how would behave the method in practice. In particular, LinLog seems more prone

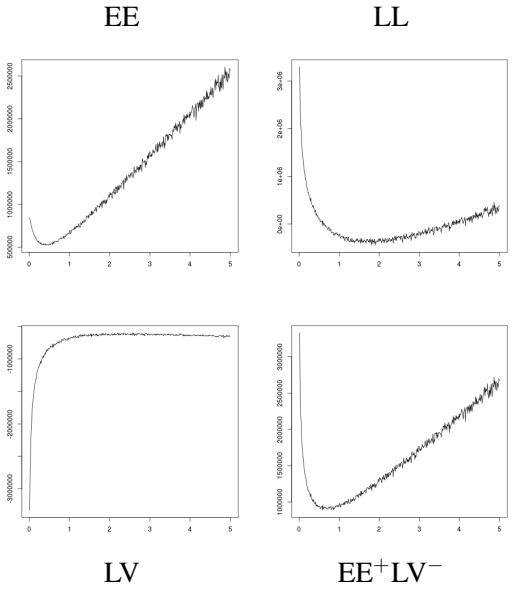

Figure 2: Values of the criteria against the variable σ^2.

to construct projections with larger clusters. The result for LargeVis leads to an optimal variance higher than Elastic Embedding for this choice of parameters. The corresponding optimal values for σ^2 are approximatively after graph reading,equal to 0.4 for EE, 1.8 for LL and 1.2 for LV. Note that for LV, the optimum exists and for a larger interval of values the concavity of the curve leads to a clear maximization of the log-likelihood. The fourth methods is with the first part of EE and the second part of LV with its optimum around 0.8. In future, a comparison with different settings of the parameters and an improvement of the modeling for the interactions, attractions and repulsions is required for further informations. For instance, the shape for the macro cluster can be chosen ellipsoisal as its accounts for the two spherical clusters.

4 Conclusion and perspectives

In this article, it is shown that Elastic Embedding is directly related to LargeVis with a particular choice of sparse weighting and with a Poisson mass function instead of the Bernoulli one. The method is compared from different points of view with alternative ones in order to understand further their differences. Two new approaches are proposed as a complement of the estimation of the distance in a particular sit-

uation when two clusters collapse into their centers with zero variances. In the two proposed approaches, the variances of the clusters are in stake in order to discuss the reasons why some methods lead to smaller clusters during the projection than other ones.

References

[1] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. nov, pp. 2579–2605, 2008, pagination: 27.

[2] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[3] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[4] J. Tang, J. Liu, M. Zhang, and Q. Mei, "Visualizing large-scale and high-dimensional data," ser. WWW '16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, pp. 287–297.

[5] M. A. Carreira-Perpiñan, "The elastic embedding algorithm for dimensionality reduction," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, pp. 167–174.

[6] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, ser. NIPS'02. Cambridge, MA, USA: MIT Press, 2002, pp. 857–864.

[7] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[8] E. J. Beh, "Simple correspondence analysis: A bibliographic review," *International Statistical Review*, vol. 72, no. 2, pp. 257–284, 2004.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, pp. 3111–3119.

[10] J.-P. Benzecri, *Correspondence Analysis Handbook*. Marcel Decker, 1992.

[11] R. Priam, "Symmetric generative methods and tsne: A short survey," in *IVAPP*, vol. 3, INSTICC. SciTePress, 2018, pp. 356–363.

[12] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

[13] A. Noack, "An energy model for visual graph clustering," in *Graph Drawing*, G. Liotta, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 425–436.

[14] ——, "An energy model for visual graph clustering," in *Graph Drawing*, G. Liotta, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 425–436.

[15] P. Eades, "A heuristic for graph drawing," *Congressus Numerantium*, vol. 42, pp. 149–160, 1984.

[16] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Softw. Pract. Exper.*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991.

[17] R. Davidson and D. Harel, "Drawing graphs nicely using simulated annealing," *ACM Trans. Graph.*, vol. 15, no. 4, pp. 301–331, Oct. 1996.

[18] H. Kettani and G. Ostrouchov, "On the distribution of the distance between two multivariate normally distributed points," 2007.