# Symmetric Generative Methods and tSNE: A Short Survey

Rodolphe Priam
*University of Southampton, University Road, Southampton SO17 1BJ, U.K.*

Keywords: Data Visualization, Generative Model, Latent Variables, tSNE, Survey.

Abstract: In data visualization, a family of methods is dedicated to the symmetric numerical matrices which contain the distances or similarities between high-dimensional data vectors. The method t-Distributed Stochastic Neighbor Embedding and its variants lead to competitive nonlinear embeddings which are able to reveal the natural classes. For comparisons, it is surveyed the recent probabilistic and model-based alternative methods from the literature (LargeVis, Glove, Latent Space Position Model, probabilistic Correspondence Analysis, Stochastic Block Model) for nonlinear embedding via low dimensional positions.

## 1 INTRODUCTION

In visualization of high-dimensional data, the observations in the available sample are vectorial: the rows or the columns of a numerical data matrix or table. An extensive literature exists and diverse approaches have been developped until today in this domain of research. Among the existing methods, t-Distributed Stochastic Neighbor Embedding (t-SNE or tSNE) (van der Maaten and Hinton, 2008) is a recent method which improves the previously proposed one called Stochastic Neighbor Embedding (SNE) (Hinton and Roweis, 2003). The idea of the SNE is to introduce soft-max probabilities which transform the matrix of distances defined in multidimensional scaling (MDS) (Sammon, 1969; Chen and Buja, 2009) into vectors of probabilities in order to deal with a Kullback-Liebler divergence instead of an Euclidean distance. As a dramatic improvement in comparison to former researches on distance matrices for visualization, tSNE is widely used today in various domains in order to proceed to data analysis. For instance, this method helps the researchers to improve any data processing which asks for an effective reduction or to proceed to the data analysis itself by looking at a synthetic view (Mahfouz et al., 2015; Shen et al., 2015; Delauney et al., 2016; Chen et al., 2017). Nextafter, this introduction presents further the notations, the purpose and the way it is achieved for a survey on related probabilistic methods.

**Data:** Let's have the available high-dimensional data as a set of data vectors in a space with $M$ dimensions as follows:

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

Let define $\mathcal{W} = (w_{ij})$ 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 neighbors graph such as a heatmap with for instance $w_{ij} = e^{-d_{ij}/\tau}$ for $\tau > 0$ when $d_{ij}$ is enough small, and $w_{ij} = 0$ otherwise. A weighted graph $G = (V, E, \mathcal{W})$ is defined from $V$, $E$, and $\mathcal{W}$ which stands respectively for the set of vertices ($i$ or $v_i$), edges ($e_{ij}$ or $(i, j)$), and weights $w_{ij}$. An edge $e_{ij}$ comes from a pair of vertices ($v_i$, $v_j$) in the graph of nearest neighbors. It is also denoted $\bar{E}$ for the set of pairs of vertices that are not neighbors.

**Reduced Vectors:** The purpose is to summarize $\mathcal{X}$ by finding relevant lower dimensional representations:

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

Generally $S = 2$ as the visualization appears in the two dimensional plane, even if three dimensions or even more remain possible. The paper is interested on the modeling of low dimensional positions via their pairwise distances/similarities plus bias/intercept terms. The parameterization involved is as follows,

$$\mathbf{y}_i^T \tilde{\mathbf{y}}_j + b_i + \tilde{b}_j \text{ or } \| \mathbf{y}_i - \tilde{\mathbf{y}}_j \|^2 \text{ or } \delta_{ij} = \delta(\mathbf{y}_i, \tilde{\mathbf{y}}_j). \quad (1)$$

where $\delta(.,.)$ is a function of a distance or similarity. Note that $\tilde{\mathbf{y}}_i$ and $\mathbf{y}_i$ are not always chosen equal. In the Euclidean case, it can be rewritten the latent terms as follows,

$$b_i + \tilde{b}_j + \mathbf{y}_i^T \tilde{\mathbf{y}}_j = \tilde{b}_i + \tilde{\tilde{b}}_j - \frac{1}{2}\|\mathbf{y}_i - \tilde{\mathbf{y}}_j\|^2. \quad (2)$$

This parameterization can be interpreted as describing latent variables in a low-dimensional Euclidean space, for the two forms just above. Some other



Priam, R.
Symmetric Generative Methods and tSNE: A Short Survey.
DOI: 10.5220/0006684303560363
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 3: IVAPP*, pages 356-363
ISBN: 978-989-758-289-9

Table 1: Methods presented (third row) in the survey with the modeling in stake for $\delta_{ij}$. The column named Probabilistic means that the method has or has not a generative/model-based foundation with a probabilistic density/mass function.

| Method name | Probabilistic | Modeling | Penalization | Input |
|---|---|---|---|---|
| MDS | no | Euclidean | no | $\mathcal{X}$, $\mathcal{W}$ |
| Laplacian Eigenmap | no | Inner product | no | $\mathcal{X}$, $\mathcal{W}$ |
| SNE, tSNE, ... | no | Divergence | yes (implicit) | $\mathcal{X}$, $\mathcal{W}$ |
| LargeVis | yes | Bernoulli, weights | yes (explicit) | $\mathcal{X}$, $\mathcal{W}$ |
| Glove | yes | Log-linear | no | $\mathcal{X}$ (or $\mathcal{W}$) |
| Probabilistic CA | yes | Poisson | no | $\mathcal{X}$ (or $\mathcal{W}$) |
| LSPM/LCPM | yes | GLM | no | $\mathcal{W}$ |
| SBM (re-parameterized) | yes | GLM | no | $\mathcal{W}$ |

methods for visualization such as LLE, KPCA, GPLVM are already compared in (Lee and Verleysen, 2007) for instance and are not presented in this survey paper as they seem to not follow this parameterization. The methods such as linear principal component analysis (PCA) or factor analysis (FA) with the embedding (Cunningham and Ghahramani, 2015) of the type $\| \mathbf{x}_i - \Omega \mathbf{y}_i \|^2$ with $\Omega$ a loading matrix, are not detailed neither. Methods with an embedding of the type $\| \mathbf{y}_i - \mathbf{y}_{(k)} \|^2$ where $\mathbf{y}_{(k)}$ is the position of a cluster as in Parametric Embedding (PE) (Iwata et al., 2007), or Probabilistic Latent Semantic Visualization (PLSV) (Iwata et al., 2008) are not detailed but briefly discussed at subsection 3.5. In the following sections it is considered only methods with the parameterization given in (1) or (2).

**Objective Function:** For finding suitable values of $\mathcal{Y}$ one can look for having $d_{ij}$ and $\delta_{ij}$ enough similar (large or small) according to a relevant criterion. The general form of the modeling of the whole set of methods presented in this survey extends the former MDS having for objective function, $\sum_{ij}(d_{ij} - \delta_{ij})^2$. It can be added the criterion $\sum_{ij} w_{ij}\delta_{ij}$ from Laplacian Eigenmap (Belkin and Niyogi, 2003) (see also in clustering, Spectral Clustering (von Luxburg, 2007)) as listed in Table 1. The general form is either a measure of distance or gap between functions of $d_{ij}$ and $\delta_{ij}$ for non probabilistic approaches, or either a likelihood from a probabilistic model with an independence hypothesis and the parameters depending on $\delta_{ij}$ for modeling $\mathbf{w}$, a random variable from $\mathcal{W}$. In the following, most of the methods include the construction of a vicinity graph with edges $w_{ij}$ before modeling the mapping otherwise such a graph is assumed to be already available.

**Illustration:** As an example, the dataset in (Girolami, 2001) is visualized with the data $\mathcal{X}$ and $\mathcal{W}$ where the later is for ten nearest neighbors and the weights are equal to 1 for any observed edge and to 0 otherwise. The visualization from a) CA+$\mathcal{X}$ (Benzecri, 1980), b) CA+$\mathcal{W}$, c) tSNE+$\mathcal{X}$ (van der Maaten and Hinton, 2008), (d) LargeVis+$\mathcal{W}$ (Tang et al., 2016) and (e) Kruskal's non-metric MDS+$\mathcal{X}$ are compared in Table 2. The considered indicators are the average of the Silhouettes (Rousseeuw, 1987) denoted S-Index and the Davies-Bouldin index (Davies and Bouldin, 1979) denoted DB-Index.

Table 2: Indicators for comparing the quality of projection from the five methods with the dataset of 1000 binarized images of handwritten digits with 10 classes.

| | (a) | (b) | (c) | (d) | (e) |
|---|---|---|---|---|---|
| S-Index | 0.01 | 0.43 | 0.50 | 0.51 | 0.03 |
| DB-Index | 5.82 | 1.69 | 0.97 | 1.34 | 2.29 |

If the visual map from CA+$\mathcal{W}$ is clearly better than from CA+$\mathcal{X}$, MDS performs between both. The two other methods lead to better separated frontiers for visualizing the natural classes but a graph with only binary weights is used for CA and LargeVis in this example.

The following sections present tSNE with its approximations and its variants, the recent generative methods for visualization and the perspectives.

# 2 tSNE, APPROXIMATIONS AND VARIANTS

In this section, tSNE is briefly presented and its variants for faster training or enhanced modeling.

## 2.1 t-Distributed Stochastic Neighbor Embedding

The method tSNE (van der Maaten and Hinton, 2008) minimizes the divergence between two discrete distributions. It is first defined:

$$p_{j|i} = \frac{\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2 / 2\sigma_i^2)}{\sum_{j' \neq j} \exp(-d(\mathbf{x}_i, \mathbf{x}_{j'})^2 / 2\sigma_i^2)}.$$

Each parameter $\sigma_i$ is set such as the perplexity $2^{-\sum_j p_{j|i} \log p_{j|i}}$ of the conditional distribution $P_i =$

$(p_{j|i})$ is equal to a positive value smaller than 50 given by the user. These values are critical for accentuating the frontiers between the natural classes, as observed in Spectral Clustering where alternative computations are available. In former approaches SNE (Hinton and Roweis, 2003) and NCA (Goldberger et al., 2005) the probabilities $P_i$ are directly used in the objective function. In tSNE it is considered a symmetric distribution $P$ as follows (with also $p_{i|i} = 0$):

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

A distribution $Q$ is defined with t-Student distributions (instead of the Gaussian ones in SNE/NCA), as follows (with also $q_{ii} = 0$):

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

This solves for the "crowding effect" where the projections of the classes are not well separated. The low dimensional positions $\mathbf{y}_i$ are finally found by minimizing the Kullback-Leibler divergence between the two distributions $P$ and $Q$, with the nonlinear and non-convex objective function,

$$C(\mathcal{Y}) = D_{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

## 2.2 Approximations

Several approximations have been introduced in the literature in order to accelerate the computation of the solution of tSNE -which has by default a quadratic complexity for the size $N$- while keeping nearly optimal values for $\mathcal{Y}$. In (van der Maaten, 2013; van der Maaten, 2014) a sparse approximation computes the quantities $p_{j|i}$ only for $\mathcal{N}_i$, the nearest neighbors of $\mathbf{x}_i$, and keeps them null otherwise. A N-body simulation and thus the Barnes-Hut (BH) algorithm (Appel, 1985) associated to fast nearest neighbors searches is proposed in (van der Maaten, 2014) in order to lower the complexity to only $O(N \log N)$ for the method BH-tSNE via tree-based procedures. See also (Kim et al., 2016) for a brief overview of the technical details. Interesting findings have been proposed recently in the literature in order to accelerate the training (Pezzotti et al., 2016; Parviainen, 2016; Kim et al., 2016) even further. A-tSNE (Pezzotti et al., 2017) improves the BH-tSNE algorithm by generating a relevant visual map before a full learning via the progressive visual analytics paradigm and approximated nearest neighbors training. The user can choose to improve some areas of the projection for steerability. Improvement or explaination of the training procedure by alternatives to the sequential approach introduced for tSNE can be found in (Nam et al., 2004; Yang et al., 2015). Next subsection presents the extensions of tSNE (and SNE) for improving the modeling foundations.

## 2.3 Variants, Extensions

Several methods propose to manage the case where several maps or several datasets are modeled. This is mainly treated in the literature via weighted sums for the probabilities $q_{ij}$. More precisely, they are called multiple maps (van der Maaten and Hinton, 2012; van der Maaten et al., 2012; Zhang et al., 2013; Xu et al., 2014), multiple view maps (Xie et al., 2011), hierarchical maps (Lee et al., 2015; Pezzotti et al., 2016). Variants of tSNE deal with time series and temporal data (Rauber et al., 2016) or graph layout (Kruiger et al., 2017) with a repulsive term as an additional penalization.

Other methods aim at improving tSNE (or SNE) by changing the divergence or the function in the softmax of $Q$ or eventually $P$. The Heavy-tailed Symmetric Stochastic Neighbor Embedding (HSSNE) (Yang et al., 2009) is a generalization of tSNE. Instead of the t-Student distribution it considers any other heavy-tailed distribution or any monotonically decreasing function. In (van der Maaten, 2009) it is proposed the t-Student distribution with $\nu$ degrees of freedom to optimize jointly with $\mathcal{Y}$. For a spherical embedding, the Euclidean distance is replaced by an inner product, hence a von Mises-Fisher (vMF) distribution is considered in vMF-SNE (Wang and Wang, 2016). Spherical embeddings are also met in two interesting alternative variants (Lunga and Ersoy, 2013; Lu et al., 2016).

tSNE and SNE minimize Kullback-Leibler divergences w.r.t. $\mathcal{Y}$, hence alternative divergences (Basseville, 2013) are possible for better robustness. A weighted mixture of two KL divergences is preferred in the method Neighborhood retrieval and visualization (NeRV) (Venna et al., 2010). In (Lee et al., 2013), it is proposed a different mixture of KL divergences, a scaled version of the generalized Jensen-Shannon divergence. In (Lee et al., 2015) it is proposed an additional improvement via multi-scale similarities. In (Yang et al., 2014) it is proposed the weighted symmetric stochastic neighbor embedding (ws-SNE) with a connection between several divergences ($\beta-$, $\gamma-$ , $\alpha-$ and Rényi-divergences). In (Narayan et al., 2015) it is presented a variant named AB-SNE with a $\alpha$-$\beta$ divergence. In (Bunte et al., 2012) it is developped a systematic comparison of many divergence measures and shown that no divergence is really better for simulated noises, the best one needs to be chosen according to each dataset. Considering models of diver-

gence with a generative setting (Dikmen et al., 2015) for the selection of the optimal extra parameters of these divergences has also been proposed in (Amid et al., 2015).

Nextafter, the presented models have an embedding of the positions $\mathbf{y}_i$ in their parameters, with different forms of objective functions.

# 3 PROBABILISTIC MODELS

In this section, the methods are defined via a probabilistic and model-based foundation. They are generally dedicated to the visualization/reduction of a graph which is constructed from vectorial data.

## 3.1 Probabilistic CA

Correspondence Analysis (CA) (Benzecri, 1980; Lebart et al., 1998) is a matricial method which is extensively studied for visualizing the rows and the columns of contingency tables. To perform CA on a two-way table, the correspondence matrix is $\mathbf{F} = \mathbf{X}/x_{..}$ while $x_{..}$ stands for its grand total, $\mathbf{r} = (r_1, \ldots, r_N)$ for the row margins, and $\mathbf{c} = (c_1, \ldots, c_M)$ for the column margins. A low-rank approximation leads to $\widehat{\mathbf{F}} = \widehat{\mathbf{X}}/x_{..}$ where $\widehat{\mathbf{X}}$ is the approximation of $\mathbf{X}$ which can be rewritten according to a reconstruction formula from the eigen vectors/values of a particular matrix. By rewritting elementwise (Beh, 2004) this matricial approximation, this leads to a Poisson approximation, a probabilistic version of CA:

$$x_{ij} \sim \mathcal{P}\left(x_{..} r_i c_j \left(1 + \mathbf{y}_i^T \tilde{\mathbf{y}}_j\right)\right),$$

or from the approximation[1] at the first order of the exponential function,

$$x_{ij} \sim \mathcal{P}\left(x_{..} e^{\mathbf{y}_i^T \tilde{\mathbf{y}}_j + \log r_i + \log c_j}\right).$$

This also explains why CA can be seen as not fully linear but this suggests also that from a graph matrix with cells proportional to $w_{ij}$ (or $p_{ij}$ from tSNE) instead of $x_{ij}$, a better visualization is expected.

## 3.2 Glove

Glove (Pennington et al., 2014) is based on a log-bilinear regression model in order to learn a vector space representation of words. The weighted regression for this model can be written as follows:

$$\begin{aligned} C(\mathcal{Y}) &= \sum_{i,j} h(x_{ij})(\mathbf{y}_i^T \tilde{\mathbf{y}}_j + b_i + \tilde{b}_j - \log x_{ij})^2 \\ &= \sum_{i,j} h(x_{ij})(\check{\mathbf{y}}_i^T \check{\tilde{\mathbf{y}}}_j - \log x_{ij})^2 . \end{aligned}$$

[1] It has also been proposed in the literature to alter the soft-max with alternative polynomial expressions but not for visualization.

The function $h(.)$ removes noisy cells in the criterion, it is equal to $(x_{ij}/100)^{3/4}$ for $x_{ij} < 100$ and equal to 1 otherwise. For the second line above, it is also denoted $\check{\mathbf{y}}_i = (b_i, 1, \mathbf{y}_i^T)^T$ and $\check{\tilde{\mathbf{y}}}_i = (1, \tilde{b}_i, \tilde{\mathbf{y}}_i^T)^T$ such as it is recognized a weighted factorization of the matrix with cell values $\log x_{ij}$, except that a component of the reductions is constrained to the value 1. This leads to the approximation,

$$x_{ij} \approx e^{\mathbf{y}_i^T \tilde{\mathbf{y}}_j + b_i + \tilde{b}_j},$$

such as Glove can be seen as a weighted version of CA, solved via a constrained factorization: a variant of MDS or Isomap (Tenenbaum et al., 2000) with particular weights. The original paper (Pennington et al., 2014) explains how to construct the matrix $(x_{ij})$ from raw textual data but any symmetric matrix $(w_{ij})$ may be used for visualization. In (Hashimoto et al., 2016) it is introduced a fully generative model, based on a negative binomial distribution *NegBin*, such that:

$$x_{ij} \sim NegBin\left(\theta, \frac{\theta}{\theta + e^{-||\mathbf{y}_i - \tilde{\mathbf{y}}_j||^2 + b_i + \tilde{b}_j}}\right),$$

where $\theta$ controls the contribution of large $x_{ij}$ as an alternative to the function $h(.)$.

## 3.3 Latent Space Position Models

In data analysis, the models named *latent space position models* (LSPM) are based on a parameterization of the generalized linear models as seen in (Hoff et al., 2002) in the binary case. In contrast to the other methods presented herein, the determistic parameters $\mathbf{y}_i$, $b_i$ and $\tilde{b}_j$ are replaced by random variables with same notation in the current subsection. In the *latent cluster position models* (LCPM) the positions $\mathbf{y}_i$ are modeled with a Gaussian mixture as a prior for their clustering (Handcock et al., 2007).

For LCPM, let denote $\mathbf{v}$ a $p \times N \times N$ array of covariates with $\mathbf{v}_{ij}$ a $p$-dimensional vector of covariates for $(i, j)$, $\beta$ the $p$-dimensional vector of covariate coefficients, $\mathbf{y}_i$ the $S$-dimensional position vector of $i$, $b = (b_1, \cdots, b_N)$ the vector of sender effects, $\tilde{b} = (\tilde{b}_1, \cdots, \tilde{b}_N)$ the vector of receiver effects, and $h(.)$ a link function. This leads to the likelihood function of the observed network $\mathbf{w} = (w_{ij})$ as follows:

$$f(\mathbf{w}|\theta, \mathbf{v}) = \prod_{i,j} f(w_{ij}|h^{-1}(\eta_{ij})).$$

For a complete bayesian model, $b_i$ and $\tilde{b}_j$ are Gaussian with means 0 and respective variances $\sigma_b^2$ and $\sigma_{\tilde{b}}^2$. It is written:

$$\eta_{ij} = \mathbf{v}_{ij}^\top \beta + \delta(\mathbf{y}_i, \mathbf{y}_j) + b_i + \tilde{b}_j .$$

The clustering is modeled through a mixture model with $G$ components where each one is a $S$-dimensional Gaussian with mean $\mu_k$, spherical variance with parameter $\sigma_k^2$ and mixing probability $\lambda_k$,

$$\mathbf{y}_i \sim \sum_{g=1}^{G} \lambda_k \mathcal{G}_S(\mu_k, \sigma_k^2 I_d).$$

Alternative distributions such as a t-Student one seem not have been tested yet for this model. For binary graphes, a Bernoulli distribution is usually considered for the modeling with $h(.)$ a logit function. The estimation of $\mathcal{Y}$ appears difficult because of the prior and posterior distributions which are highly nonlinear. Latent space models are able to find the natural classes and to select their number according to the best fit. They can facilitate principled visualization in a probabilistic setting. Theory on these models can be found in (Rastelli et al., 2016) and faster inference in (Raftery et al., 2012) even if the application of these models may be confined to moderated sizes of graphs for the current available implementations. The next method can be seen as a regularized LSPM with an efficient implementation and no bayesian priors.

## 3.4 LargeVis

LargeVis (Tang et al., 2016) introduces a probabilistic parametric model dedicated to the visualization of a nearest neighbors graph. The model is defined for not only the set of the nearest neighbors but also all the furthest ones which are forced into a negative interaction. It is closely related to the previous approach LINE (Tang et al., 2015b) (see also (Cao et al., 2015) for an alternative matrix-based learning and (Tang et al., 2015a) for a semi-supervised variant) which is non generative. A likelihood for the graph $G$ can then be written as follows,

$$L(\mathcal{Y}) = \prod_{(i,j)\in E} p(e_{ij}=1)^{w_{ij}} \prod_{(i,j)\in \bar{E}} (1-p(e_{ij}=1))^{\gamma}.$$

Where,

$$P(e_{ij}=1) = g(||\mathbf{y}_i - \mathbf{y}_j||^2),$$

stands for the probability that a edge $e_{ij}$ exists between $v_i$ and $v_j$. For the function $g(.)$, it can be chosen,

$$g(\tau) = 1/(1+a\tau) \text{ or } g(\tau) = 1/(1+exp(\tau)).$$

while $\gamma$ is a common weight for the non neighbor vectices. For a spherical version with $g(\mathbf{y}_i^T \mathbf{y}_j)$, a vMF distribution function induces an embedding over a sphere as in subsection 2.3. LargeVis recalls a Bernoulli distribution which has been altered for improving its properties of separability of the clusters by adding the weigths and the penalization. Concerning this modeling, two remarks are proposed :

- Let's have $\bar{w} = \sum_{(a,b)\in E} w_{ab}$, $\tilde{p}_{ij} = w_{ij}/\bar{w}$, $\tilde{\gamma} = \gamma/\bar{w}$, $\tilde{q}_{ij} = p(e_{ij}=1)$. By rewriting $L(.)$, it is obtained a new function to minimize w.r.t. $\mathcal{Y}$,

$$\begin{aligned} & \ell_\gamma(\mathcal{Y}) \\ = & -\frac{\log L(\mathcal{Y})}{\bar{w}} + \sum_{(i,j)\in E} \tilde{p}_{ij} \log \tilde{p}_{ij} \\ = & \sum_{(i,j)\in E} \tilde{p}_{ij} \log \frac{\tilde{p}_{ij}}{\tilde{q}_{ij}} - \tilde{\gamma} \sum_{(i,j)\in \bar{E}} \log(1-\tilde{q}_{ij}). \end{aligned}$$

  Hence, it is recognized a criterion in two parts. The term on the left side is roughly similar to the criterion of tSNE as a divergence but without $(\tilde{q}_{ij})$ normalized to sum to 1 and without t-Student distributions. The other term on the right side is for a penalization. They insure the local and global projections respectively.

- It seems appealing to try to improve other distributions such as a Poisson one which may replace the first term,

$$P(e_{ij}=w_{ij}) = p(e_{ij}=1)^{w_{ij}},$$

  to get eventually an expressive quantity for the non observed edges with $P(e_{ij}=0)$. A probabilistic interpretation -when $w_{ij}$ is an integer- remains the product between the likelihood of the observed edges with the likelihood of the non observed edges with a weighting for regulating the importance of each one:

$$L_{\mathcal{P}}(\mathcal{Y}) = \prod_{(i,j)\in E} \frac{(\tilde{\delta}_{ij})^{w_{ij}} e^{-\tilde{\delta}_{ij}}}{w_{ij}!} \prod_{(i,j)\in \bar{E}} e^{-\gamma \tilde{\delta}_{ij}}.$$

  Here $\tilde{\delta}$ is the result from a function of the quantity $\delta_{ij}$ such as the exponential one. In this alternative likelihood, the weighting is modeled explicitely and the model is fully generative. The term to the left with a Poisson mass distribution for $E$ recalls LSPM without bayesian priors. When in $\tilde{\delta}_{ij}$ an exponential function is chosen, the term to the right for $\bar{E}$ recalls the penalization in Elastic Embedding (Carreira-Perpiñan, 2010), except the weighting. Additional weighting is via the parameters with respectively, $\tilde{\delta}_{ij} = e^{\delta_{ij}+\log\alpha}$ for $E$ and $\tilde{\delta}_{ij} = e^{\delta_{ij}+\log w_{ij}}$ for $\bar{E}$. With $\alpha > 0$, this parameterization may lead to a weighting similar to Elastic Embedding for the non observed edges.

## 3.5 Stochastic Block Model

As presented in (Matias and Robin, 2014; Daudin et al., 2008) the stochastic block model (SBM) is defined for a random graph on a set $V = \{1,\ldots,N\}$ of $N$ nodes ($i$ or $v_i$). Let's have $\mathbf{z} = \{z_1,\ldots,z_N\}$ stands for $N$ independent and identically distributed (i.i.d.)

discrete hidden random variables with possible values in $\{1,\dots,K\}$. Let's have $f_\gamma(\cdot;\gamma_{z_i z_j})$ a conditional density function or mass distribution with parameter $\gamma=(\gamma_{k\ell})_{1\le k,\ell\le K}$. The variables $w_{ij}$ are random, i.i.d. conditionally to $(z_i,z_j)$ and aggregated into the random variable $\mathbf{w}=(w_{ij})_{(i,j)\in E}$ with a given distribution. SBM has for conditional likelihood the following expression, where $z_i$ and $z_j$ needs to be integrated out by adding their distribution,

$$\begin{array}{rcl} f(\mathbf{w}|\mathbf{z}) & = & \prod_{(i,j)\in E} f(w_{ij}|z_i,z_j) \\ & = & \prod_{(i,j)\in E} f_\gamma(w_{ij}|\gamma_{z_i z_j}) \,. \end{array}$$

The links between the edges and the structure of the network are sometimes explained by covariates: $\mathbf{v}_i$ at the node level as in (Tallberg, 2004) via a multinomial probit model for the membership of the vertices or $\mathbf{v}_{ij}$ at the edge level as in (Mariadassou et al., 2010) via a regression term within the expectations. This modeling concerns mainly the clustering part of the stochastic block model which needs to be reparameterized for inducing a nonlinear visualization (if a posteriori methods such as Parametric Embedding (Iwata et al., 2007) are not used). For visualization purposes with this model, further parameters can be embedded. This results into adding the $N$ latent variables $\mathbf{y}_i$ via $\delta(\mathbf{y}_i,\mathbf{y}_j)$ or its corresponding cluster version $\delta(\mathbf{y}_{(k)},\mathbf{y}_{(\ell)})$ with eventual bias terms ($b_i$ and $\tilde{b}_j$ or the cluster versions $b_{(k)}$ and $\tilde{b}_{(\ell)}$). In the binary case when $g(.)$ is the sigmoidal function, $f_\gamma$ can be written with a Bernoulli mass distribution function with parameters,

$$\gamma_{k\ell} = g(\parallel \mathbf{y}_{(k)} - \mathbf{y}_{(\ell)} \parallel^2) \,.$$

This re-parameterization of SBM recall LSPM but with a different clustering framework as the mixture model is not a prior but directly introduced in the data modeling.

A limitation of the approaches above may be seen in diagonal co-clustering (Tjhi and Chen, 2006): this suggests that the quantities $\gamma_{z_i z_j}$ could not be fully free parameters. The extra parameters for the visualization needs to be added in the posterior probabilities for instance. This leads in the variational EM for the inference of SBM (with common parameters or not) to consider the probability that a datum belongs to a cluster as one of the following expression:

$$\begin{array}{rcll} Q_{y_i}(z_i=k;\gamma) & \propto & e^{-\|\mathbf{y}_i-\mathbf{y}_{(k)}\|^2} & \text{as in PLSV}, \\ Q_{y_i}(z_i=k;\gamma) & = & \sum_{k'} h_{kk'}\tau_{ik'} & \text{as in SOM}. \end{array}$$

For the later case at the bottom, such algorithm introduces the quantities $\tau_{ik'}$ as the free parameters and $h_{kk'}$ as a smoothing matrix from the neighboor nodes in the Kohonen's network or self-organizing maps (SOM) (Kohonen, 1997).

# 4 CONCLUSION AND PERSPECTIVES

In this survey, it is proposed an unified overview of the literature on data visualization with tSNE and with the recent alternative symmetric generative methods[2] depending on bivariate latent positions. Several links between these methods are explained for helping the comparisons of their objective functions. These comparisons suggest eventual variants of several existing methods such as: CA estimated approximatively via Glove for large matrices, LSPM or SBM regularized via a probabilistic penalization for the non observed edges, or the visualization with the mixture models extended to symmetric matrices via SBM for a symmetric self-organizing map for instance, as future appealing perspectives.

# ACKNOWLEDGEMENTS

The author would like to thanks the reviewers for the valuable comments.

# REFERENCES

Amid, E., Dikmen, O., , and Oja, E. (2015). Optimizing the information retrieval trade-off in data visualization using α-divergence. *ArXiv e-prints*.

Appel, A. W. (1985). An efficient program for many-body simulation. *SIAM Journal on Scientific and Statistical Computing*, 6(1):85–103.

Basseville, M. (2013). Review: Divergence measures for statistical data processing-An annotated bibliography. *Signal Process.*, 93(4):621–633.

Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review*, 72(2):257–284.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Compututation*, 15(6):1373–1396.

Benzecri, J. P. (1980). *L'analyse des données tome 1 et 2 : l'analyse des correspondances*. Paris:Dunod.

Bunte, K., Haase, S., Biehl, M., and Villmann, T. (2012). Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45.

[2]The presented methods have very different training algorithms according to their targeted domain (datamining, data analysis, data visualization or machine learning) and the size of the datasets their are experimented with. Most of these methods can ask for diverse estimation algorithms such as bayesian inference or online gradient optimization with efficient data structure, efficient memory management, in order to infer in the faster way $\mathcal{Y}$.

Cao, S., Lu, W., and Xu, Q. (2015). GraRep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM'15, pages 891–900.

Carreira-Perpiñan, M. A. (2010). The elastic embedding algorithm for dimensionality reduction. In *Proceedings of the 27th International Conference on Machine Learning*, ICML'10, pages 167–174.

Chen, L. and Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219.

Chen, V., Paisley, J., and Lu, X. (2017). Revealing common disease mechanisms shared by tumors of different tissues of origin through semantic representation of genomic alterations and topic modeling. *BMC Genomics*, 18(Suppl 2):105.

Cunningham, J. P. and Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16:2859–2900.

Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Delauney, C., Baskiotis, N., and Guigue, V. (2016). Trajectory bayesian indexing: The airport ground traffic case. In *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1047–1052.

Dikmen, O., Yang, Z., and Oja, E. (2015). Learning the information divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7):1442–1454.

Girolami, M. (2001). The topographic organization and visualization of binary data using multivariate-bernoulli latent variable models. *IEEE Transactions on Neural Networks*, 12(6):1367–1374.

Goldberger, J., Hinton, G. E., Roweis, S. T., and Salakhutdinov, R. R. (2005). Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007). Model-based clustering for social networks. *JRSS-A*, 170(2), 301–354.

Hashimoto, T. B., Alvarez-Melis, D., and Jaakkola, T. S. (2016). Word embeddings as metric recovery in semantic spaces. *TACL*, 4:273–286.

Hinton, G. E. and Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15*, pages 857–864. MIT Press.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, Theory and Methods*, 97(460).

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T. L., and Tenenbaum, J. B. (2007). Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556.

Iwata, T., Yamada, T., and Ueda, N. (2008). Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'08, pages 363–371.

Kim, M., Choi, M., Lee, S., Tang, J., Park, H., and Choo, J. (2016). PixelSNE: Visualizing Fast with Just Enough Precision via Pixel-Aligned Stochastic Neighbor Embedding. *ArXiv e-prints*.

Kohonen, T. (1997). *Self-organizing maps*. Springer.

Kruiger, J. F., Rauber, P. E., Martins, R. M., Kerren, A., Kobourov, S., and Telea, A. C. (2017). Graph Layouts by t-SNE. *Comput. Graph. Forum (Proc. of EuroVis)*, 36(3):283–294.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*. Springer.

Lee, J. A., Peluffo-Ordonez, D. H., and Verleysen, M. (2015). Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261.

Lee, J. A., Renard, E., Bernard, G., Dupont, P., and Verleysen, M. (2013). Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer, 1st edition.

Lu, Y., Yang, Z., and Corander, J. (2016). Doubly Stochastic Neighbor Embedding on Spheres. *ArXiv e-prints*.

Lunga, D. and Ersoy, O. (2013). Spherical stochastic neighbor embedding of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2):857–871.

Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M. J., and Lelieveldt, B. P. (2015). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods*, 73:79–89.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Ann. Appl. Stat.*, 4(2):715–742.

Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings*, 47:55–74.

Nam, K., Je, H., and Choi, S. (2004). Fast stochastic neighbor embedding: a trust-region algorithm. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 1, page 123–128.

Narayan, K., Punjani, A., and Abbeel, P. (2015). Alpha-beta divergences discover micro and macro structures in data. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML'15, pages 796–804.

Parviainen, E. (2016). A graph-based N-body approximation with application to stochastic neighbor embedding. *Neural Networks*, 75:1–11.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, pages 1532–1543.

Pezzotti, N., Höllt, T., Lelieveldt, B. P. F., Eisemann, E., and Vilanova, A. (2016). Hierarchical stochastic neighbor embedding. *Comput. Graph. Forum (Proc. of EuroVis)*, 35(3):21–30.

Pezzotti, N., Lelieveldt, B. P. F., van der Maaten, L, Höllt, T.,Eisemann, E., and Vilanova, A. (2017). Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1739–1752.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

Rastelli, R., Friel, N., and Raftery, A. E. (2016). Properties of latent variable network models. *Network Science*, 4(4):407–432.

Rauber, P. E., Falcão, A. X., and Telea, A. C. (2016). Visualizing time-dependent data using dynamic t-SNE. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: Short Papers*, EuroVis'16, pages 73–77.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

Shen, F., Shen, C., Shi, Q., van den Hengel, A., Tang, Z., and Shen, H. T. (2015). Hashing on nonlinear manifolds. *IEEE Transactions on Image Processing*, 24(6):1839–1851.

Tallberg, C. (2004). A bayesian approach to modeling stochastic blockstructures with covariates. *The Journal of Mathematical Sociology*, 29(1):1–23.

Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016). Visualizing large-scale and high-dimensional data. In *WWW'16*.

Tang, J., Qu, M., and Mei, Q. (2015a). PTE: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'15, pages 1165–1174.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015b). LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW'15, pages 1067–1077.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Tjhi, W.-C. and Chen, L. (2006). A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recogn. Lett.*, 27(3):151–159.

van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, volume 5, pages 384–391.

van der Maaten, L. (2013). Barnes-Hut-SNE. In *Proceedings of International Conference on Learning Representations*.

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, 15:3221–3245.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

van der Maaten, L. and Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55.

van der Maaten, L., Schmidtlein, S., and Mahecha, M. D. (2012). Analyzing floristic inventories with multiple maps. *Ecological Informatics*, 9:1–10.

Venna, J., Peltonen, J., Nybo, K., Aidos, H., and Kaski, S. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wang, M. and Wang, D. (2016). vMF-SNE: Embedding for spherical data. In *ICASSP'16*, Shanghai, China.

Xie, B., Mu, Y., Tao, D., and Huang, K. (2011). m-SNE: Multiview stochastic neighbor embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1088–1096.

Xu, W., Jiang, X., Hu, X., and Li, G. (2014). Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. *BMC Med. Genomics*, 7:1–9.

Yang, Z., King, I., Xu, Z., and Oja, E. (2009). Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 22*, pages 2169–2177. Curran Associates.

Yang, Z., Peltonen, J., and Kaski, S. (2014). Optimization equivalence of divergences improves neighbor embedding. In *Proceedings of the 31st International Conference on on Machine Learning*, ICML'14, pages 460–468.

Yang, Z., Peltonen, J., and Kaski, S. (2015). Majorization-minimization for manifold embedding. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS'15)*, volume 38, pages 1088-1097.

Zhang, L., Zhang, L., Tao, D., and Huang, X. (2013). A modified stochastic neighbor embedding for multi-feature dimension reduction of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 83:30–39.