A LARGE-DIMENSIONAL ANALYSIS OF SYMMETRIC SNE

Charles Séjourné, Romain Couillet, Pierre Comon

LargeDATA chair, GIPSA-lab, University Grenoble-Alpes

ABSTRACT

Stochastic Neighbour Embedding methods (SNE, t-SNE) aim at finding a faithful low-dimensional representation of a high-dimensional dataset. Despite their popularity, being solution to a non-convex optimization, the behavior of these tools is not well understood. This work provides first answers by leveraging a large dimensional statistics approach, where the number n and dimension p of the large-dimensional data are of the same magnitude. We derive and study the canonical equation verified by the critical points of this non-convex optimization problem. The study notably reveals that, in a simple setup, the achievable SNE solutions correspond to a subset of those critical points. In particular, when the clusters composing the dataset are balanced in size, these solutions are symmetrical and assume closed-form expressions.

As a major conclusion, the analysis rigorously proves a long-standing heuristic statement on the "proper normalization" of the symmetric SNE: out of two natural normalization choices, only the claimed proper one leads to non-trivial solutions.

Index Terms— dimensionality reduction, high-dimension, non-convex, random matrix theory, machine learning

1. INTRODUCTION

Stochastic Neighbor Embeddings (SNE) [1] and its variant t-SNE[2] are dimensionality reduction methods used for visualization purposes (usually in 2 or 3 dimensions) as a pre-analysis tool for data scientists to get a grasp on their (usually numerous and large dimensional) data. The low-dimensional representations achieved by those algorithms stem from a non-convex optimization problem, traditionally solved via gradient-descent. As such, it is far from trivial to predict the resulting solution, let alone to interpret the associated visualization. Concretely, the following natural concern arises: considering 3 equidistant vectors in $\mathbb{R}^p$ with $p > 2$,[1] how would SNE or t-SNE represent those points when projected in $\mathbb{R}$? The equidistant property cannot hold in $\mathbb{R}$ and it is unclear which representation SNE will retain. The subject of the present article is precisely to cast a first light on this riddle.

Despite the underlying non-convex optimization from which SNE solutions arise, recent tractable analyses of SNE algorithms emerged in the machine learning literature. Using dynamical system techniques, [3] highlights a "shrinkage" property of the low-dimensional representations: under mild assumptions, within each cluster, the data point projections converge to each other as the gradient descent progresses. This is however not sufficient to ensure consistent clustering as different clusters are not guaranteed not to merge. The analysis of [4] then builds on [3] and shows that the cluster centroids in the low-dimensional representation space remain separated; combined with [3], this ensures that the representations form well-separated clusters.

While those articles establish the consistency of the SNE methods, they do not provide an accurate view on the solution landscape: how many solutions exist? With which associated cost? Is gradient descent guaranteed to fall in a global minimum? The present work addresses some of these aspects. This however demands to break the problems of non-linearity and non-convexity of the cost function to be minimized. To this end, we use a large n, p data assumption and exploit large dimensional asymptotics. For simplicity of exposition and for the sake of retrieving insightful results, the n data follow a mixture of k large dimensional Dirac masses in $\mathbb{R}^p$.

Our main results may be summarized as follows:

- we obtain a *limiting* vector equation as, n, p grow large at the same rate, satisfied by the critical points of the underlying SNE optimization problem, and analyze their stability;
- the limiting equation appears to *slightly but fundamentally* differ depending on the specific normalization of the symmetric SNE method initially introduced by [2]: as a result, we reach a first rigorous argument in favor for the SNE normalization effectively used today in practice;
- the observed low-dimensional outputs are shown to be severely biased by the number of components of each class, thereby disrupting the original SNE objective;
- in a balanced class-size setting, the limiting vector equation assumes a closed-form expression, which reveals the surprising presence of a *continuum of saddle points* and, as a result, a very slow convergence of SNE, when compared to faster convergence in unbalanced scenarios.

The remainder of the article presents the symmetric SNE algorithm and the large-dimensional setup used for our analysis (Section 2), before introducing our main results under the form of the limiting vector equation satisfied by the symmetric SNE critical points (Section 3) which is analyzed in a simple setting and turned into a closed-form expression. For readability, in what follows, "SSNE" will refer to the symmetric SNE [2].

Couillet's work is partially supported by the ANR-MIAI LargeDATA chair at University Grenoble-Alpes, and the HUAWEI LarDist project.

[1] Say for instance 3 canonical basis vectors of $\mathbb{R}^p$.

2. PROBLEM FORMULATION

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be input high-dimensional data vectors, and $\boldsymbol{X} = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$. The goal of SNE is to associate with each $x_i \in \mathbb{R}^p$ a vector $y_i \in \mathbb{R}^d$, for $d < p$ (in general $d \in \{2, 3\}$, but we shall subsequently also consider $d = 1$ for simplicity), such that $\boldsymbol{Y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^{n \times d}$ preserves as much of the significant structure of $\boldsymbol{X}$ in $\mathbb{R}^d$ as possible.

2.1. The SSNE algorithm

The idea of SSNE as devised by [2] is to build two joint probability distributions P and $Q \in \mathbb{R}^{n \times n}$ for the data points $\boldsymbol{X}$ and their (yet to be found) low-dimensional representations $\boldsymbol{Y}$, respectively. Starting from some initialization point, the representations $\boldsymbol{Y}$ are then iteratively tuned in order for the density Q to accurately approximate P.

Precisely, for some positive and non-increasing functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the density matrix Q is defined as

$$Q_{ij} = \begin{cases} \frac{G_{ij}}{\sum_{k \neq l} G_{kl}} & , \ i \neq j \\ 0 & , \ i = j \end{cases}, \quad G_{ij} = g(\|y_i - y_j\|^2)$$

while P is alternatively defined, for $F_{ij} = f(\frac{1}{p}\|x_i - x_j\|^2)$, as

$$P_{ij} \underset{i \neq j}{=} \begin{cases} P_{\mathrm{MN};ij} & = \frac{F_{ij}}{\sum_{k \neq l} F_{kl}} \\ P_{\mathrm{RMNS};ij} & = \frac{1}{2n}\left(\frac{F_{ij}}{\sum_{l \neq i} F_{il}} + \frac{F_{ij}}{\sum_{l \neq j} F_{lj}}\right) \end{cases}, \ P_{ii} = 0$$

where MN stands for "matrix normalization" and RMNS for "row-matrix normalization and symmetrization", discussed in [1] as two normalization procedures for P. The authors in [1] empirically claim that the RMNS solution yields better visualizations, as it is likely more robust to outliers. One objective of the article is to theoretically support this statement. In the original SNE algorithm [2], both f and g are taken to be $f(t) = g(t) = \exp(-t/2)$, thereby mimicking a Gaussian distribution for both low and high dimensional data representations. The t-SNE approach rather relies on a Student-t modification of the method in which $g(t) = 1/(1+t)$, the justification of which is also heuristic rather than based on theoretical supports.

For simplicity of exposition, in the sequel, we restrict ourselves to the case where $g(t) = \exp(-t/2)$.

The low-dimensional representation matrix $\boldsymbol{Y}$ is then chosen in such a way to minimize the Kullback-Leibler divergence between P and Q

$$C(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{KL}(P||Q) = \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right). \tag{1}$$

Not being a convex problem in $\boldsymbol{Y}$, a local minimum is then determined using a mere gradient descent procedure. A pleasant advantage of this formulation is that, for $g(t) = \exp(-t/2)$ as considered here, the gradient at running point $\boldsymbol{Y}$ is simply defined as [2]

$$\frac{\partial C}{\partial y_i} = 4\sum_{j=1}^{n}(P_{ij} - Q_{ij})(y_i - y_j), \quad i = 1, \ldots, n. \tag{2}$$

The solutions of the SNE problem, which correspond to the local minima of $C(\boldsymbol{X}, \boldsymbol{Y})$, are therefore associated with some of the solutions of the set $\{\partial C/\partial y_i = 0, \ i = 1, \ldots, n\}$, i.e., the critical points of C. The objective of the article is to precisely characterize these critical points, under a statistically convenient setting. To this end, it is convenient to provide an expression of the Hessian matrix $H(\boldsymbol{Y})$ at any given point $\boldsymbol{Y}$. Specifically, we find that

$$H(\boldsymbol{Y}) = 4\Big[(Q - P) \otimes I_d - \frac{2}{\gamma_p} S - \frac{4}{\gamma_p^2} T \mathbb{1}_n \mathbb{1}_n^T T^T - \mathcal{D}\Big(((Q - P) \otimes I_d)\mathbb{1}_{nd} - \frac{2}{\gamma_p} S \mathbb{1}_{nd}\Big)\Big]$$

where $\gamma_p = \sum_{k \neq l} G_{kl}$ is the normalization constant ensuring that Q is a probability distribution, $\mathbb{1}_n$ is the vector of ones of size n, $\otimes$ is the Kronecker product, $\mathcal{D}(v)$ is the diagonal matrix with vector v on the diagonal, and matrices S and T of size $nd \times nd$ and $nd \times n$ respectively defined block-wise as

$$\{S_{i \times d + l, j \times d + m}\}_{0 \leq l, m \leq d-1} = G_{ij}(y_j - y_i)(y_j - y_i)^T$$
$$\{T_{i \times d + l, j}\}_{0 \leq l \leq d-1} = G_{ij}(y_i - y_j).$$

2.2. The large dimensional setting

As the purpose of SNE is to visualize large dimensional data of $\mathbb{R}^p$ in the low dimensional space $\mathbb{R}^d$, the article considers the setting where p, n are arbitrarily large and of the same order of magnitude, while d is fixed, i.e., $n, p \to \infty$ with $p/n \to \gamma > 0$. We further assume the x_i's drawn from

$$\mathcal{L} = \sum_{\ell=1}^{k} c_\ell \delta_{\mu_\ell}, \quad c_1, \ldots, c_k > 0, \quad \sum_{\ell=1}^{k} c_\ell = 1$$

that is, a mixture of k masses located at $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$. Moreover, as $p \to \infty$, $\|\mu_\ell\| = O(1)$ for each ℓ.[2] We may write

$$\boldsymbol{X} = J\boldsymbol{\mu}^T, \ \boldsymbol{\mu} = [\mu_1, \ldots, \mu_k] \in \mathbb{R}^{p \times k}, \ J = [j_1, \ldots, j_k] \in \mathbb{N}^{n \times k}$$

where $[j_\ell]_i = \delta_{\{x_i = \mu_\ell\}}$ is the indicator vector of class ℓ. In particular, denoting $n_\ell \equiv |\{x_i = \mu_\ell\}|$, by the law of large numbers, $n_\ell/n \to c_\ell$ almost surely.

As shall be seen, this simple setup already allows one to retrieve interesting insights on the SSNE algorithm behavior.

3. MAIN RESULTS

A few notations need be defined before introducing our main results. Since all identical $x_i \in \mathbb{R}^p$ (say $x_i = \mu_\ell$) are expected to have the same representation $y_i \in \mathbb{R}^d$, $\boldsymbol{Y}$ must be of the form $\boldsymbol{Y} = J\tilde{\boldsymbol{Y}}$, where $\tilde{\boldsymbol{Y}} \in \mathbb{R}^{k \times d}$. We similarly denote $\widetilde{Q} \in \mathbb{R}^{k \times k}$ the matrix such that $Q = J\widetilde{Q}J^T - \frac{f(0)}{\gamma_p} I_n$, where the diagonal centering is needed since Q has a null diagonal.

3.1. Critical points

With these notations and the assumptions of the previous section at hand, we have the following result.

[2] This condition is crucial and mimics the non-trivial setting of machine learning classification of e.g., [5] where $x_i \sim \mathcal{N}(\mu_a, \sigma^2 I_p)$ with $\|\mu_a\|, \sigma^2 = O(1)$, but here in the limit where $\sigma^2 = 0$.

Theorem 1 (Asymptotic null gradient condition)**.** *As $n, p \to \infty$, every solution $\tilde{\boldsymbol{Y}}$ of (2) satisfies*

$$\|L\tilde{\boldsymbol{Y}}\| \to 0, \quad L \equiv \mathcal{D}(\mathbb{1}_{n_\bullet})A\mathcal{D}(\mathbb{1}_{n_\bullet}) - \mathcal{D}(\mathbb{1}_{n_\bullet})\mathcal{D}(A\mathbb{1}_{n_\bullet}) \quad (3)$$

where

$$A = \mathbb{1}_k\mathbb{1}_k^T - h(M) - n(n-1)\widetilde{Q}$$
$$h(M) = \frac{f'(0)}{2pmf(0)} \times \begin{cases} \frac{2}{n}\mathbb{1}_{n_\bullet}^T M \mathbb{1}_{n_\bullet}\mathbb{1}_k\mathbb{1}_k^T - 2nM, & P = P_{MN} \\ MD_{n_\bullet}\mathbb{1}_k\mathbb{1}_k^T + \mathbb{1}_k\mathbb{1}_k^T D_{n_\bullet}M - 2nM, & P = P_{RMNS} \end{cases}$$

for $\mathbb{1}_{n_\bullet} = [n_1, \ldots, n_k]^T$, $M = \{\|\mu_a - \mu_b\|^2\}_{a,b=1}^k \in \mathbb{R}^{k\times k}$ and $\mathcal{D}_v$ is the diagonal matrix with vector v on the diagonal.

Sketch of Proof. Exploiting $\|\mu_\ell\| = O(1)$ so that $\frac{1}{p}\|x_i - x_j\|^2 \to 0$ for all i, j, following the ideas of [5], a Taylor expansion on the entries of P is performed in the large n, p limit which, along with the symmetry of the problem, reduces $\{\frac{\partial C}{\partial y_i} = 0,\ 1 \le i \le n\}$ from an n- to a k-dimensional equation with f only involved through $f(0)$ and $f'(0)$. □

It is interesting to observe that Theorem 1 holds identically under both MN and RMNS normalizations, but for the expression of the function $h(M)$ which contains the structural information about the data statistics $\mu_1, \ldots, \mu_k$.

According to Theorem 1, for large n, p, determining the critical points of the SSNE problem boils down to solving the k-dimensional equation $L\tilde{\boldsymbol{Y}} = 0$ in which the non-linear function f appearing in P is now reduced to its derivatives $f(0)$ and $f'(0)$. Unlike f though, since d is small, the non-linear function g, intervening in Q_{ij}, remains encapsulated within $\widetilde{Q}$ and cannot be simplified. Equation $L\tilde{\boldsymbol{Y}} = 0$ thus remains difficult to analyze due to the involved non-linear relation between $\boldsymbol{Y}$ and $\widetilde{Q}$.

3.2. Simplified setting

For readability and ease of interpretation, we consider from now on the case where $k = 3$, $d = 1$, and $\mu_a = me_a$, for $e_a \in \mathbb{R}^p$ the a-th canonical vector of $\mathbb{R}^p$ and $m > 0$. By a common shift-invariance of the entries of $\tilde{\boldsymbol{Y}}$ solution to $L\tilde{\boldsymbol{Y}} = 0$, $\tilde{\boldsymbol{Y}}$ may be parametrized only with (say) $\tilde{y}_1 - \tilde{y}_2$ and $\tilde{y}_1 - \tilde{y}_3$.

Under this simplified setting, the MN approach can be shown to provide *trivial* solutions:

Proposition 1. *In the MN case, any $\tilde{\boldsymbol{Y}} \in \mathbb{R}^3$ such that $\tilde{y}_i = \tilde{y}_j$ for at least one combination of $i \neq j \in \{1, 2, 3\}$ is a critical point of $C(\tilde{\boldsymbol{Y}})$.*

Sketch of Proof. The result follows from direct calculus, fundamentally exploiting the (here quite unfavorable) symmetrical nature of P_{MN} (unlike for P_{RMNS}, in this setting, $P_{\text{MN},ij}$ is constant for all $i \neq j$ and all $i = j$). □

The MN critical points raised in Proposition 1 are highly undesirable as they "merge" distinct clusters of $\mathbb{R}^p$ together in $\mathbb{R}^d$ according to the cardinality of the classes. Figure 2, which reports the landscape of $\|L\tilde{\boldsymbol{Y}}\|$ under both MN and RMNS settings in a unbalanced cluster-size scenario, effectively pinpoints those critical points in the MN setting and shows that a symmetrical pair of them are indeed local minima. This is quite unlike the RMNS method which does not lead to trivial solutions, suggesting that RMNS is always a preferred method. This is quite at odds with the seemingly "more natural" MN approach to minimizing the Kullbach-Liebler divergence in (1); as pointed out in the proof sketch of Proposition 1, RMNS here paradoxically benefits from its being "less symmetrical" than MN.

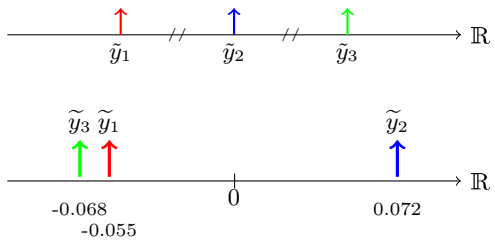

Fig. 1: Visualization of SSNE solutions $\tilde{\boldsymbol{Y}} = [\tilde{y}_1, \tilde{y}_2, \tilde{y}_3]^T$, for $d = 1$, $k = 3$, $p = 1\,024$ and $\mu_a = me_a$ (e_a the a-th canonical vector of $\mathbb{R}^p$) with $m = 5$. **(Top)** Any of the 6 symmetrical solutions in the balanced setting ($n_1 = n_2 = n_3 = 500$). **(Bottom)** Most typical solution for the unbalanced setting with $n_1 = 300$, $n_2 = 700$, $n_3 = 500$.

Figure 2 provides extra empirical information. In particular, it identifies a total of seven MN but only five RMNS critical points. Excluding the trivial $\tilde{y}_1 = \tilde{y}_2 = \tilde{y}_3$ solution, the other critical points in fact reduce to symmetrical pairs (changing the signs of all $\tilde{y}_i$'s in a solution brings the other). Among the two RMNS pairs, only one *asymptotically* corresponds to a minimum (green bullets in the figure), which we empirically observed to be associated with the scenario where, in the visualization of dimension $d = 1$, the central cluster $\tilde{y}_i$ is the one with lowest c_i value (see Figure 2-(bottom)); this local minimum is here global.

A complete understanding of all critical points remains nonetheless tedious, unless the clusters have balanced sizes $c_1 = \ldots = c_k$, which the next section is devoted to explore.

3.3. Balanced mixture ($c_1 = \ldots = c_k$) for $k = 3$, $d = 1$

In the setting of the previous section, let now $c_1 = c_2 = c_3$. It is easy to observe that MN and RMNS now coincide; indeed, we have in this case $\mathbb{1}_{n_\bullet} = (n/k)\mathbb{1}_k$ so that

$$\frac{2}{n}\mathbb{1}_{n_\bullet}^T M \mathbb{1}_{n_\bullet}\mathbb{1}_k\mathbb{1}_k^T = MD_{n_\bullet}\mathbb{1}_k\mathbb{1}_k^T + \mathbb{1}_k\mathbb{1}_k^T D_{n_\bullet}M.$$

What is less obvious though is that this balanced cluster size scenario triggers a strongly degenerate behavior of the solutions $\tilde{\boldsymbol{Y}}$.

Theorem 2. *Under the above assumptions, for both MN and RMNS, the set of points $\tilde{\boldsymbol{Y}}$ satisfying $L\tilde{\boldsymbol{Y}} = 0$ is the union of the singleton $\{\tilde{y}_1 = \tilde{y}_2 = \tilde{y}_3\}$ and of the points parametrized by the ellipse*

$$\tilde{y}_1 - \tilde{y}_3 = \sqrt{\frac{-2f'(0)}{15pf(0)}}m\left(3\cos t - \sqrt{3}\sin t\right),$$
$$\tilde{y}_1 - \tilde{y}_2 = \sqrt{\frac{-2f'(0)}{15pf(0)}}m\left(3\cos t + \sqrt{3}\sin t\right), \quad t \in \mathbb{R}.$$

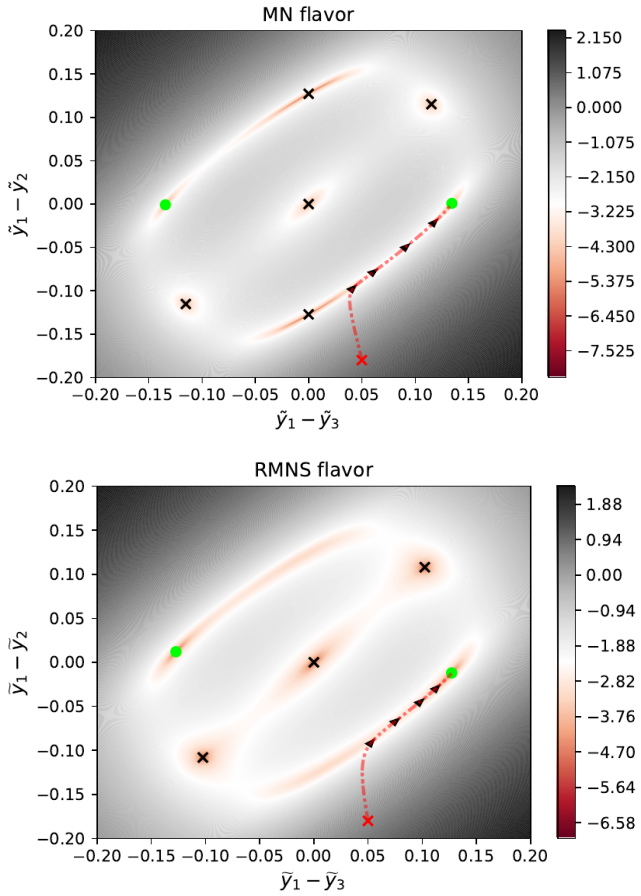

Fig. 2: Value of $\log\left(\|L\tilde{Y}\|^2\right)$ for MN versus RMNS; $n_1 = 300, n_2 = 700, n_3 = 500, m = 5, p = 1024$. Local minima in green bullets. Saddle points in black crosses. Gradient descent sample path in red dotted line.

As such, Theorem 2 states that there *asymptotically* is an infinity of critical points. Figure 3 corroborates the result and numerically suggests, through the local evaluation of the Hessian $H(\boldsymbol{Y})$, that the ellipse of critical points is the successive union of saddle-point and local-minimum contiguous regions.

The presence of contiguous critical points has a severe impact on convergence, as depicted by the red paths in Figures 2–3 for which we use the same initialization points and the same gradient descent hyperparameters. While the isolated minima in the unbalanced cluster-size scenario is a strong attractor (Figure 2) and is ultimately reached at the end of the descent, in the balanced cluster-size scenario gradient descent first reaches an arbitrary point of the limiting ellipse (Figure 3) before slowly moving towards one of the symmetrical solutions where $|\tilde{y}_1 - \tilde{y}_2| = |\tilde{y}_2 - \tilde{y}_3|$ (or any recombination of indices); these symmetrical solutions are symbolized in red diamonds in Figure 3 and correspond to the scenario depicted in Figure 1-(top)). After reaching the ellipse though, as $n, p \to \infty$, gradient descent slows down significantly, to the point of barely progressing as soon as n, p are of the order of a few thousands: therefore, in this

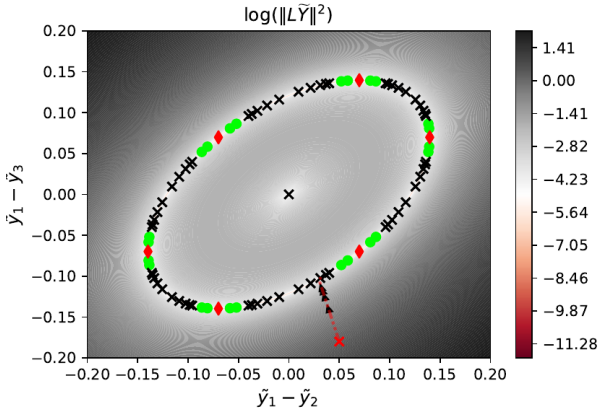

Fig. 3: Value of $\log\left(\|L\tilde{Y}\|^2\right)$ for $n_1 = n_2 = n_3 = 500, m = 5, p = 1024$. Local minima in green bullets. Saddle points in black crosses. Gradient descent sample path in red dots. Symmetrical solutions achieved at finite n, p values in red diamonds.

scenario, SSNE, even in the RMNS flavor, dramatically fails.

4. CONCLUDING REMARKS

The article demonstrates that, despite the non-convex nature of the stochastic neighborhood embeddings algorithm, large dimensional statistics is able to capture its dominant behavior. This behavior is largely non-trivial and quite counter-intuitive: (i) the most symmetrical scenario suffers from degeneracy and failure to converge efficiently, (ii) the stable $\mathbb{R}^d$ visual outputs do not appropriately translate the distances between clusters in the original space $\mathbb{R}^p$ but are impacted by cluster sizes, an undesirable feature, and (iii) the advocated RMNS approach, preferred over the nonetheless "more natural" MN flavor, only relies in a favorable asymmetrical behavior of the distribution matrix P of the affinities $f(\frac{1}{p}\|x_i - x_j\|^2)$ of the data in ambient space.

The work yet only scratches the surface of a more complete understanding of SNE methods: many questions remain open as to the behavior under more realistic noisy data inputs, the extension to the utmost popular (and even more asymmetrical) t-SNE method, and as to "what real data visualizations really tell?".

5. REFERENCES

[1] Geoffrey E Hinton and Sam T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds., pp. 857–864. MIT Press, 2003.

[2] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[3] George C. Linderman and Stefan Steinerberger, "Clustering with t-sne, provably," *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 2, pp. 313–332, 2019.

[4] Sanjeev Arora, Wei Hu, and Pravesh K Kothari, "An analysis of the t-sne algorithm for data visualization," *arXiv preprint arXiv:1803.01768*, 2018.

[5] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.