Manifold Analysis for High-Dimensional Socio-Environmental Surveys

Charles Dupont and Debraj Roy[0000-0003-1963-0056]

Faculty of Science, Informatics Institute, University of Amsterdam,
Science Park 904, 1090 XH, Amsterdam, the Netherlands.
{c.a.dupont,d.roy}@uva.nl

Abstract. Recent studies on anthropogenic climate change demonstrate a disproportionate effect on agriculture in the Global South and North. Questionnaires have become a common tool to capture the impact of climatic shocks on household agricultural income and consequently on farmers' adaptation strategies. These questionnaires are high-dimensional and contain data on several aspects of an individual (household) such as spatial and demographic characteristics, socio-economic conditions, farming practices, adaptation choices, and constraints. The extraction of insights from these high-dimensional datasets is far from trivial. Standard tools such as Principal Component Analysis, Factor Analysis, and Regression models are routinely used in such analysis, but they either rely on a pairwise correlation matrix, assume specific (conditional) probability distributions, or assume that the survey data lies in a linear subspace. Recent advances in manifold learning techniques have demonstrated better detection of different behavioural regimes from surveys. This paper uses Bangladesh Climate Change Adaptation Survey data to compare three non-linear manifold techniques: Fisher Information Non-Parametric Embedding (FINE), Diffusion Maps and t-SNE. Using a simulation framework, we show that FINE appears to consistently outperform the other methods except for questionnaires with high multipartite information. Although not limited by the need to impose a grouping scheme on data, t-SNE and Diffusion Maps require hyperparameter tuning and thus more computational effort, unlike FINE which is non-parametric. Finally, we demonstrate FINE's ability to detect adaptation regimes and corresponding key drivers from high-dimensional data.

Keywords: Survey Analysis · Climate Change Adaptation · Fisher Information · t-SNE · Diffusion Maps

1 Introduction

Climate change is one of the significant global challenges of the 21st century and floods are the costliest climate-induced hazard. Rapid urbanization and climate change exacerbate flood risks worldwide, undermining humanity's aspirations to achieve sustainable development goals (SDG) [12]. Current global warming trends and their adverse impacts such as floods represent a complex problem,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_3

which cannot be understood independently of its socioeconomic, political, and cultural contexts. In particular, the impact of climate change on farmers and their livelihoods is at a critical juncture and adaptation is key for embracing best practices as new technologies and pathways to sustainability emerge. As the amount of available data pertaining to farmers' adaptation strategies has increased, so has the need for robust computational methods to improve the facility with which we can extract insights from high-dimensional survey data. Although standard methods such as Principal Component Analysis, Factor Analysis or Regression models are routinely used and have been effective to some degree, they either rely on a pairwise correlation matrix, assume specific (conditional) probability distributions or that the high-dimensional survey data lies in a linear subspace [6]. Recent advances in manifold learning techniques have shown great promise in terms of improved detection of behavioural regimes and other key non-linear features from survey data [2].

In this paper, we compare three non-linear manifold learning techniques: t-SNE [14], Diffusion Maps [4], and Fisher Information Non Parametric Embedding (FINE) [1]. We start by extending prior work [8] done with a simulation framework which allows for the generation of synthetic questionnaires. Because the underlying one-dimensional statistical manifolds are known, we are able to quantify how well each algorithm is able to recover the structure of the simulated data. Next, we apply the various methods to the Bangladesh Climate Change Adaptation Survey [10], which contains rich data regarding aspects of individual households such as spatial information, socio-economic and demographic indicators, farming practices, adaptation choices and constraints to adaptations. This allows us to investigate whether behavioural regimes of adaptation can be extracted, and more broadly to better understand each method's utility and relative trade-offs for the analysis of high-dimensional, real world questionnaires. Although all three methods yield comparable results, we uncover key differences and relative advantages that are important to take into consideration. By virtue of being non-parametric, FINE benefits from decreased computational efforts in contrast to t-SNE and Diffusion Maps which require hyperparameter tuning. Although FINE typically outperforms the other two methods, its performance degrades when there is high interdependence between survey items, and we identify a cutoff point beyond which adding more features does not result in increasing the differential entropy of pairwise distances between observations in the resulting embedding. FINE requires the researcher to impose a grouping scheme on observations, which may not always be intuitive, while t-SNE and Diffusion Maps allow clusters (groups of similar observations) to emerge more naturally since no prior structure is assumed. Nonetheless, FINE is shown to be particularly successful in the extraction of adaptation regimes. Lastly, FINE allows one to use as much data as possible since missing feature values can simply be ignored, whereas t-SNE and Diffusion Maps can be significantly impacted by imputed or missing values, which may require removing incomplete observations.

The structure of the rest of the paper is as follows. Section 2 provides an overview of the algorithms studied in this work, as well as the simulation framework, climate change adaptation questionnaire, and experiments that are carried out. Section 3 presents key results obtained for the various experiments. Lastly, Section 4 discusses our findings as well as future directions of research.

2 Methods

2.1 Dimension Reduction Algorithms

t-SNE t-Distributed Stochastic Neighbour Embedding (t-SNE) was first introduced by Laurens van der Maaten and Geoffrey Hinton [14], and is based on prior work on Stochastic Neighbour Embedding [9]. Key steps are presented and summarized in Supporting Information (SI), Algorithm 1. First, a probability distribution over pairs of data points in the original feature space is constructed such that the similarity of some data point $\mathbf{x}_j$ to data point $\mathbf{x}_i$ is defined as

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

We can interpret this quantity as the conditional probability of selecting $\mathbf{x}_j$ as a neighbour of $\mathbf{x}_i$. $p_{i|i} = 0$ since a data point cannot be its own neighbour. σ_i denotes the variance of the Gaussian distribution centered around $\mathbf{x}_i$. It is tuned for each $\mathbf{x}_i$ separately such that the resulting conditional probability distribution P_i over all other datapoints $\mathbf{x}_{j \neq i}$ yields a perplexity value specified by the user, calculated as perplexity$(P_i) = 2^{H(P_i)}$, where $H(P_i)$ is the Shannon entropy [14]. Because typically $p_{j|i} \neq p_{i|j}$, we define the joint distribution $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2N}$, where N denotes the total number of observations in the dataset. The next step is to construct another probability distribution over the data in a lower-dimensional space with the aim of minimizing the Kullback-Leibler (KL) divergence between the previous probability distribution and this newly constructed one, thus preserving similarities between data points in the original space. The joint probabilities for data points in this lower dimensional map are given by

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

which is a heavy-tailed Student t-distribution [14]. KL divergence is minimized iteratively using gradient descent by updating vectors $\mathbf{y}_i$ at each step.

Diffusion Maps Diffusion Maps is a method introduced by Coifman and Lafron which takes inspiration from the processes of heat diffusion and random walks [4]. Intuitively, if we were to take a random walk over observations in a dataset, starting at some random point, we would be more likely to travel to a nearby, similar point than to one that is much further away. The Diffusion Maps algorithm leverages this idea in order to estimate the connectivity k between pairs of data points using a Gaussian kernel as follows: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\epsilon}\right)$,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_3

where ϵ is some normalization parameter. Subsequently, a diffusion process is constructed using the transition matrix of a Markov Chain M on the data set, which allows us to map diffusion distances to a lower-dimensional space. Key parameters are t, used to construct the t-step transition matrix M^t, and additional normalization parameter α. SI Algorithm 2 summarizes the important steps of this process.

FINE The Fisher Information Non Parametric Embedding (FINE) algorithm was developed by Carter et al. and works by constructing a statistical manifold upon which lives a family of probability distributions (estimated from some dataset) for which we can compute inter-distances [1]. This algorithm was further developed and applied to questionnaire data by Har-Shemesh et al. [8]. Algorithm 1 summarizes key steps. First, respondents to the questionnaire are divided into K groups. For each of these groups, a probability distribution is constructed over the set of all possible responses I (each element being a string, e.g. "ABDC"). By considering the square roots of these probabilities, we can regard each probability distribution as a point on the unit hypersphere and compute distances between these points using the arc length. With this distance matrix, non-linear dimension reduction is achieved by applying classical Multidimensional Scaling (MDS), which is another non-linear technique for visualizing similarities between observations in a dataset [11]. Questionnaire items are assumed to be independent such that probabilities may be factorized.

Algorithm 1: FINE (for questionnaire data)

```
Data: $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
Input: dimension $d$, choice of grouping scheme
Result: $E$, a lower-dimensional representation of the data
begin
    Divide observations into $K$ groups using some grouping scheme;
    for $k = 1, 2, \ldots, K$ do
        Estimate (square root) probabilities $\xi^{(k)}$ for responses in $k^{th}$
          group;
    end
    for $j, k = 1, 2, \ldots, K$ do
        Compute $M_{ij} = \cos^{-1}\left(\sum_I \xi_I^{(j)} \xi_I^{(k)}\right) \rightarrow$ arc length on the unit
          hypersphere;
    end
    Construct embedding $E = \mathrm{MDS}(M, d)$;
end
```

2.2 Simulation Framework

Framework Description The authors of [8] propose a simulation framework for generating questionnaire responses in a controlled way. This allows us to compare the embeddings generated by the three algorithms to a "ground-truth" embedding. This is done by parameterising angles ϕ_i as

$$\phi_i^{\kappa}(t) = \begin{cases} \frac{\pi}{2}\sin^2(m\pi t), & i = \kappa \\ \frac{\pi}{2}t, & i \neq \kappa \end{cases}, \tag{1}$$

where κ allows us to choose which angle is proportional to the squared sine term, and $t \in [0,1]$ is the unique parameter of this family of probability distributions. Furthermore, m controls the non-linearity of the family. There are $N-1$ angles for a questionnaire with N possible distinct responses, and we compute the square root probabilities as follows:

$$\begin{aligned} \xi_1 &= \cos(\phi_1) \\ \xi_2 &= \sin(\phi_1)\cos(\phi_2) \\ \xi_3 &= \sin(\phi_1)\sin(\phi_2)\cos(\phi_3) \\ &\vdots \\ \xi_{N-1} &= \sin(\phi_1)\ldots\sin(\phi_{N-2})\cos(\phi_{N-1}) \\ \xi_N &= \sin(\phi_1)\ldots\sin(\phi_{N-2})\sin(\phi_{N-1}) \end{aligned} \tag{2}$$

For some choice K, which denotes the total number of groups (see Algorithm 1), we draw K values uniformly on the curve given by Equation (1). Then, we compute probabilities $p_I^{(k)} = (\xi_I^{(k)})^2$ and randomly generate a number of questionnaire responses for each group $k = 1, 2, \ldots, K$ using these probabilities.

Experiments We wish to compare the embeddings generated by t-SNE, Diffusion Maps, and FINE for various simulated questionnaire responses. Similarly to [8], we generate responses for $\kappa \in \{1, 2, N-1\}$, and $K \in \{20, 50\}$. We keep $m = 3$ fixed as well as the number of questions ($N_Q = 8$) and the number of possible answers for each question ($N_A = 3$), yielding $N = 3^8 = 6561$. For each of the 6 possible combinations of parameters κ and K, there is a unique theoretical embedding and 30 questionnaires are simulated. When $K = 20$ we generate 25 responses per group, and when $K = 50$ we generate 50 responses per group. Then, for each set of 30 questionnaires, we apply all three non-linear dimension reduction algorithms.

In order to evaluate the quality of the generated embeddings, we apply the Procrustes algorithm [7], which can stretch, rotate or reflect the generated embeddings so that they match up with the theoretical embedding as closely as possible. Once this is done, we compute the Pearson correlation coefficient between the coordinates of each generated embedding and those of the theoretical embedding. Note that the theoretical embedding is determined via application of the MDS algorithm using arc length distances between the exact probability distributions calculated using Equation (2).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_3

Parameter Tuning FINE does not require any parameterization, although a grouping scheme must be provided, which in this case is defined by the simulation framework. On the other hand, both t-SNE and Diffusion Maps require some parameter tuning. For t-SNE, we perform a grid search over the following parameters: perplexity $\in \{1, 2, 5,\ 10\}$, learning rate $\eta \in \{10, 50, 100, 200\}$, distance metric $\in$ {weighted hamming (with/without one-hot encoding), cosine (with one-hot encoding)}. The maximum number of steps T and momentum $\alpha(t)$ are fixed at 1000 and 0.8 respectively. For Diffusion Maps, we perform a grid search over: $\epsilon \in \{0.5, 1.0, 1.5, 2.0\}$, $t \in \{0, 0.5, 1, 5\}$, distance metric $\in$ {weighted hamming (with/without one-hot encoding), cosine (with one-hot encoding)}. We fix $\alpha = \frac{1}{2}$. See [3] for a review of one-hot encoding, and note that the weighted hamming distance is simply the number of positions that two strings differ, each positional contribution (1 if different, 0 if identical) being weighted by the reciprocal of the number of possible values at that position.

2.3 Bangladesh Climate Change Adaptation Survey

The non-linear manifold learning algorithms of interest are applied to a questionnaire dataset pertaining to the economics of adaptation to climate change in Bangladesh with the aim of identifying different regimes of behaviour and adaptation in response to climate change.

Dataset Description Data collection was carried out in 2012 amongst 827 households in Bangladesh in 40 different communities [10]. This survey is a follow-up to a first round of data collection, which was studied in detail in [5]. Some households have frequently missing response fields, so we retain 805 households having responded to at least 30% of survey questions. Each of the 40 distinct communities has a unique combination of district, “upazila”, and union codes, where upazilas are sub-units of districts, and unions are even smaller administrative units. Households additionally possess one of 7 distinct codes corresponding to different agro-ecological zones.

Handpicked Features We construct a set of handpicked features that we expect to be important for detecting adaptation strategies based on existing literature. Specifically, we keep track of household income and expenditure, what occupations are held by household members, total monetary loss due to climatic and personal shocks, what actions were taken in response, social capital, collective action, constraints to adaptations, what adaptations were implemented, and finally what community groups household members are a part of as well as associated benefits. This set of 95 features is summarized in SI Table 8. Summary statistics for various features are also provided in other SI tables. Note that we discretise continuous features into at most 5 bins using the Bayesian Blocks dynamic programming method, first introduced by Scargle [13]. Additional details regarding this method are available in the SI document.

DOI: 10.1007/978-3-031-36027-5_3

Experiments We apply FINE to the set of handpicked features, using communities as our grouping scheme for individual household observations. Additionally, we examine the impact of how much a particular feature varies across communities on the embedding produced by FINE as follows. For each handpicked feature, we compute the KL divergence of that feature's values for each pair of communities. We record the median, and after producing an embedding using FINE, we compute the differential entropy of the distribution of pairwise distances between communities. Finally, we apply t-SNE and Diffusion Maps to the set of handpicked features in order to see if any clusters naturally emerge.

3 Results

3.1 Simulation Framework Results

Table 1 displays the best hyperparameter combinations for t-SNE and Diffusion Maps. For t-SNE, lower perplexity values typically perform better, as does the weighted hamming distance. For Diffusion Maps, using the cosine distance metric (with one-hot encoding) yields optimal performance for all (κ, K) pairs. However, choices for ϵ and t seem to be more delicate and dependent on the κ, K values. Overall, for both t-SNE and Diffusion Maps, not all hyperparameter combinations are found to yield good performance, emphasizing the importance of hyperparameter tuning.

Table 1: Summary of best hyperparameters for t-SNE and Diffusion Maps

(κ, K)	t-SNE	Diffusion Maps
$(\kappa = 1, K = 20)$	perplexity $= 2$, $\eta = 50$	$\epsilon = 1.5$, $t = 0.5$
$(\kappa = 1, K = 50)$	perplexity $= 5$, $\eta = 200$	$\epsilon = 1.5$, $t = 0.5$
$(\kappa = 2, K = 20)$	perplexity $= 2$, $\eta = 50$	$\epsilon = 0.5$, $t = 0.5$
$(\kappa = 2, K = 50)$	perplexity $= 10$, $\eta = 200$	$\epsilon = 0.5$, $t = 0.5$
$(\kappa = 6560, K = 20)$	perplexity $= 2$, $\eta = 50$	$\epsilon = 2.0$, $t = 0.5$
$(\kappa = 6560, K = 50)$	perplexity $= 1$, $\eta = 50$	$\epsilon = 2.0$, $t = 0.0$

Figure 1 displays the distribution of performance (correlation with theoretical embedding) of each algorithm for all 30 questionnaires and each (κ, K) combination using best-performing hyperparameters. FINE significantly outperforms the other algorithms in all cases and with lower variance in performance across the 30 questionnaires, except when $\kappa = 1$ and $K = 20$ where Diffusion Maps performs similarly. t-SNE consistently performs worse, and is significantly more sensitive to which of the 30 questionnaires is being analyzed (as evidenced by the high variance in performance). Overall, all algorithms achieve a mean correlation with the theoretical embedding of at least 0.87 (at a 95% confidence level).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_3

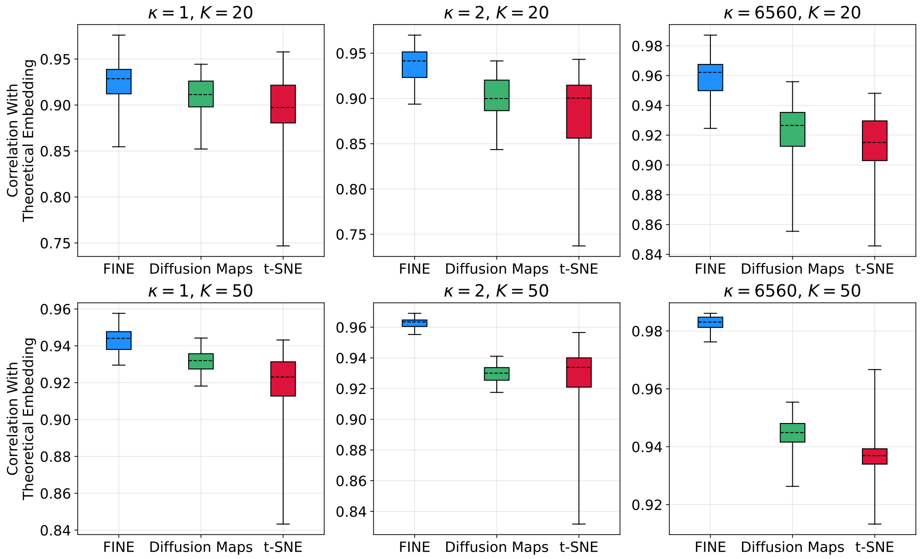

Fig. 1: Distributions of correlation coefficients with respect to the theoretical embedding for each algorithm and (κ, K) pair using best hyperparameters. Dashed lines and whiskers denote the mean, maximum and minimum values.

Figure 2 displays the theoretical embedding for each $(\kappa,\ K)$ pair, along with embeddings obtained using t-SNE, Diffusion Maps, and FINE for one sample questionnaire. Overall the embeddings produced by FINE most closely resemble the theoretical embeddings out of all three algorithms. Additionally, the underlying structure of the data is better recovered with a larger number of groups K and responses in all cases except for t-SNE when comparing $(\kappa = 1, K = 20)$ and $(\kappa = 1, K = 50)$. Multi-partite information, which measures the amount of dependence between the questions of the simulated questionnaires, is also displayed and is defined as

$$\mathrm{MI} \equiv \sum_I p_I(q_1, q_2, \ldots, q_{N_Q}) \ln \frac{p_I(q_1, q_2, \ldots, q_{N_Q})}{p_I(q_1)p_I(q_2)\ldots p_I(q_{N_Q})}. \tag{3}$$

In order to more closely investigate the dependence of the various algorithms' performance on multi-partite information, we generate an additional 300 questionnaires (each one with its own theoretical embedding and distribution of multi-partite information values), using $N_Q = 7$, $N_A = 3$, and 30 uniformly spaced κ values between 1 and $N - 1$ as well as $m \in \{1, 2, \ldots, 10\}$. We fix the number of groups at $K = 20$, and generate 50 responses per group. As always, FINE does not require any parameter tuning. For t-SNE, using Table 1 as a guide, we use perplexity $= 2$, $\eta = 50$ and a weighted hamming distance metric after one-hot encoding. For Diffusion Maps, relying on Table 1, we select $\epsilon = 0.5$ and $t = 0.5$, and use the cosine distance metric after one-hot encoding.

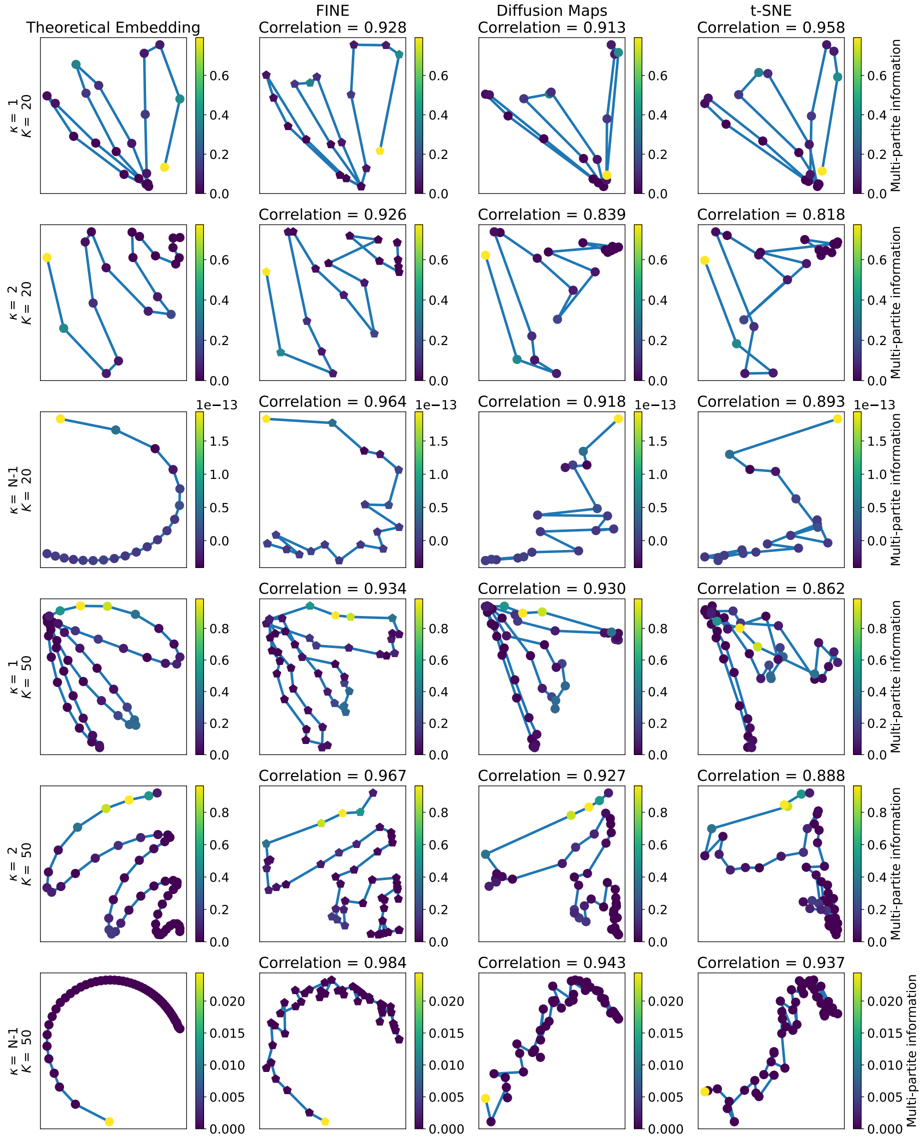

Fig. 2: Comparison of embeddings obtained with t-SNE, Diffusion Maps, and FINE with respect to the theoretical embedding using the simulation framework. t-SNE achieves comparable performance to FINE due to hyperparameter tuning, which is a departure from prior results presented in [8].

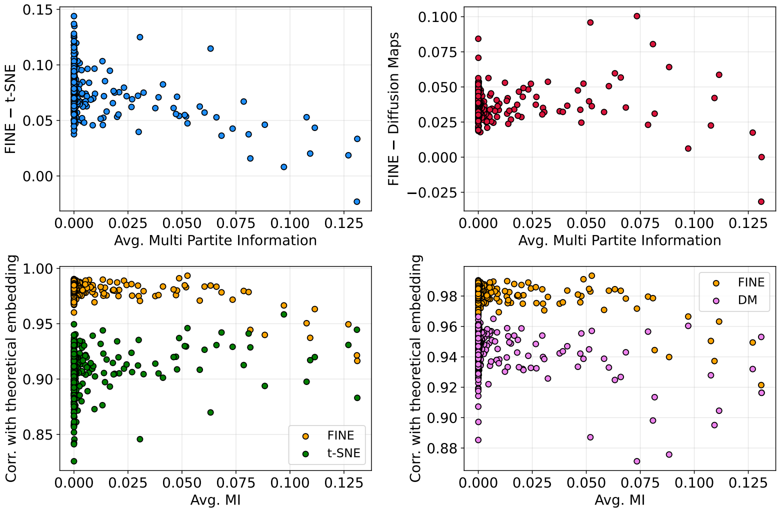

Fig. 3: (Top row) difference between FINE performance and t-SNE (left) as well as Diffusion Maps (right) as a function of multi-partite information, abbreviated MI. (Bottom row) FINE's performance begins to degrade for questionnaires with higher multi-partite information. t-SNE (left) and Diffusion Map's (right) performances are overlaid in green and pink respectively.

The top row of Figure 3 displays the differences in correlation with respect to the theoretical embedding between FINE and t-SNE as well as FINE and Diffusion Maps for 300 different values of (averaged) multi-partite information. In agreement with Figure 1, the differences are almost always positive, indicating that FINE typically outperforms the other two algorithms. However, at higher values of average multi-partite information, FINE's performance starts to worsen relative to both t-SNE and Diffusion Maps. Looking at the bottom row of Figure 3, we can tell from the yellow markers that FINE's performance decreases around MI values of 0.075. In contrast, t-SNE's performance appears to improve, while Diffusion Map's performance seems to remain the same on average. The degradation in FINE's performance may be attributable to the fact that FINE assumes independence between questions and therefore does not handle situations where there is higher interdependence between survey items as well.

3.2 Bangladesh Climate Change Adaptation Survey Results

Figure 4 illustrates the FINE embeddings obtained as we progressively add more handpicked features, starting with ones with lower median KL divergence. The top left embedding includes a single feature corresponding to monetary loss due to sea level rise with median KL divergence close to zero, which signifies that almost all communities have the same distribution for this feature. This results in a distribution of pairwise distances with very low differential entropy

– nearly all communities collapse to the same coordinate, except for community 6 which appears as an outlier due to being the only one containing a household having suffered damages due to sea level rise. Community 21 also appears as a clear outlier in subsequent embeddings. Upon investigation, we found that 65.5% of households in this community reported having to migrate due to suffering heavy losses as a result of soil and river erosion, which is significantly more than households in any other community.

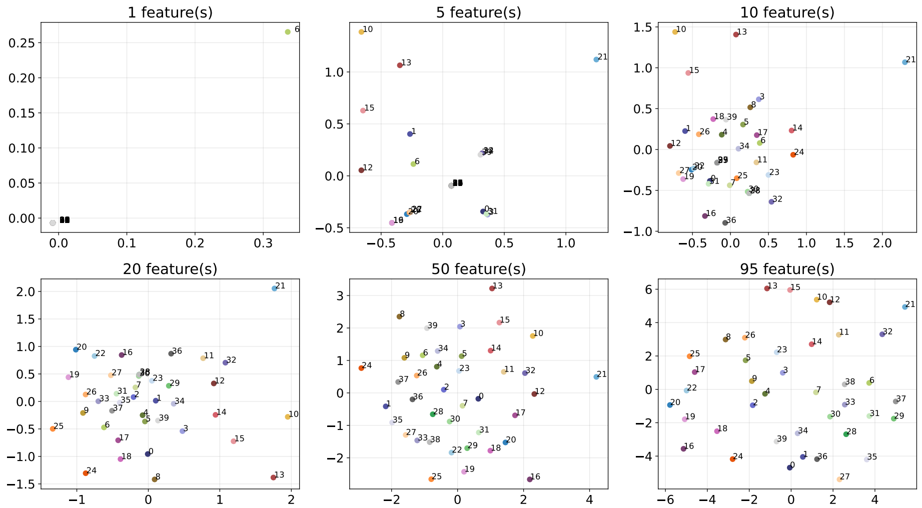

Fig. 4: FINE embeddings using an increasing number of handpicked features, added in order from lowest to highest median KL divergence.

The bottom right subplot includes all 95 handpicked features. Despite not including the agro-ecological zone in the set of handpicked features, we notice the influence of spatial characteristics on adaptation regimes quite clearly in some cases. For example, communities 29, 30, 31, and 33 all appear close together in the embedding and in fact are all located in the same agro-ecological zone. Since agro-ecological zones are defined as regions with similar climate conditions, it is perhaps unsurprising that communities in the same geographical areas would be similarly impacted by climatic shocks as well as respond in a similar fashion. However, such proximity is certainly not the only driver of adaptation. Communities 10, 12, 13, and 15 appear close together at the top of the embedding, but belong to three different agro-ecological zones. In fact, these are the only communities in which at least two households needed to sell assets in response to salinity increases. Furthermore, communities 10 and 12 had over 30% of households with at least one member needing to seek off-farm employment, which could explain their appearing especially close together. Only communities 21 and 32 also have this property, and as a result they appear quite isolated in the embedding as well

DOI: 10.1007/978-3-031-36027-5_3

(especially 21 for reasons mentioned earlier). As a last example, despite being spread over four different agro-ecological zones, at least 65% of households in communities 16-20, 22, 24 and 25 decided to change their planting dates, and notice that these communities form an elongated vertical cluster in the bottom left of the embedding. The only other community satisfying this property is community 15, which differs in other, more pronounced regards (as described earlier). Additional embeddings were generated using FINE for various feature sets, which may be found in the SI document.

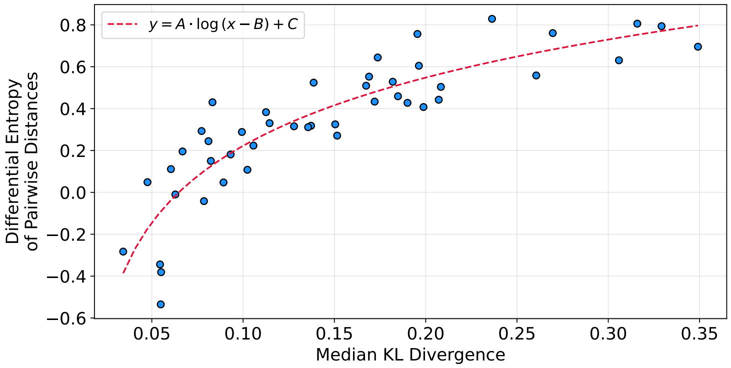

Fig. 5: Dependence of the differential entropy of pairwise distances in FINE embeddings on median KL divergence between communities for handpicked feature set. Best fit parameters: $A = 0.422$, $B = 0.014$, $C = 1.26$.

Figure 5 displays how the differential entropy of pairwise distances between communities behaves as a function of the median KL divergence for embeddings produced with one handpicked feature at a time. We observe a logarithmic trend, which seems to imply that past a certain threshold, a feature containing more information and richer differences between communities does not necessarily yield a distribution of pairwise distances with higher entropy.

We now turn our attention to the top row of Figure 6, which shows the embeddings obtained by t-SNE and Diffusion Maps for the set of handpicked features after one-hot encoding and removing any households with missing values for any of the features, leaving a total of 256 households. t-SNE uses a weighted hamming distance while Diffusion Maps relies on a cosine distance metric. Households appear closely packed together with no discernible clusters, which we find to be consistent across different runs of t-SNE and Diffusion Maps. In the bottom row, we plot community barycenters by collapsing households in the same community to their mean coordinates. The cluster of communities 16-20, 22, 24 and 25 emerges somewhat for both algorithms. However, the cluster of communities 10, 12, 13, and 15 is not clear-cut for t-SNE, and Diffusion Maps does not highlight community 21 as an outlier.

DOI: 10.1007/978-3-031-36027-5_3

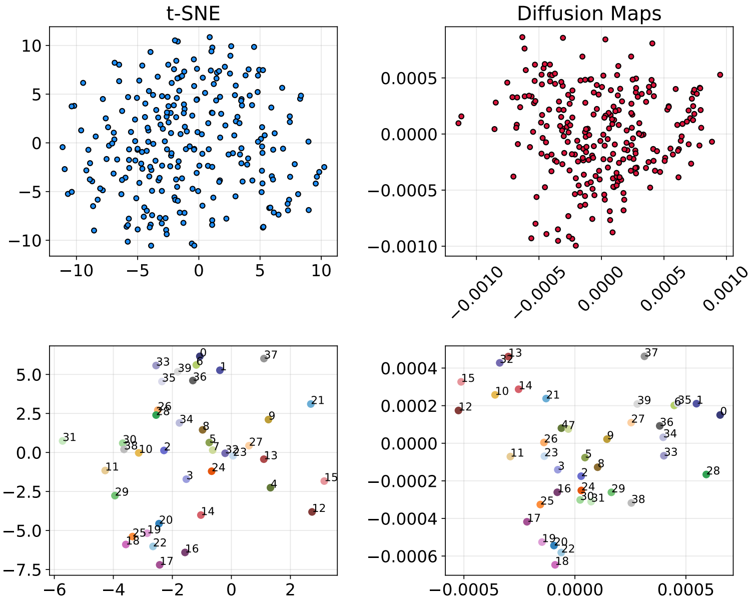

Fig. 6: (Top left) t-SNE embedding with handpicked features after one-hot encoding, using a weighted hamming distance metric. (Top right) Diffusion Maps embedding for same feature set after one-hot encoding, using cosine distance metric. (Bottom row) community barycenters for t-SNE and Diffusion maps.

Lastly, we compare pairwise distances between community coordinates for each pair of algorithms using handpicked features. Overall, we find that there is agreement across all three methods regarding the arrangement of the communities in relation to one another. Pearson correlation coefficients are found to be: 0.569 (t-SNE and Diffusion Maps), 0.514 (FINE and Diffusion Maps), and 0.318 (FINE and t-SNE). Perhaps unsurprisingly, correlation is highest between t-SNE and Diffusion Maps since the community grouping scheme was not applied to these two methods and many households were omitted due to missing feature values. An additional visualization of these correlations may be found in SI Figure 3.

4 Discussion

Experiments carried out with a simulation framework reveal that all three methods achieve comparable performance in terms of recovering the general structure of the underlying one-dimensional manifolds. This is a departure from the previous study using this framework, which underestimated the performance of t-SNE in particular due to a lack of hyperparameter tuning. FINE appears to consistently outperform the other two methods except for questionnaires with high multi-partite information. This reduction in performance may be due to

DOI: 10.1007/978-3-031-36027-5_3

the assumption that FINE makes about independence between survey items, which could potentially be relaxed for strongly-correlated survey questions. It remains to be seen how FINE responds to a wider range of MI values. Pathways to simulating questionnaires with higher MI values include investigating the dependence of multi-partite information on the parameters of the framework, or the injection of dependencies by duplicating feature columns and introducing noise. Estimating multi-partite information for real-world datasets, such as the high-dimensional survey studied in this paper, remains challenging since the product of marginals in Equation (3) quickly tends to zero.

The embeddings obtained with FINE reveal a logarithmic convergence in terms of how much dispersion of the communities can be observed in the embeddings as a function of how much feature values vary between communities. Indeed, features behaving similarly for many groups yield embeddings with strong clusters and significant overlap, whereas groups appear much more spread out for features with more variability between groups. The choice of grouping scheme is therefore non-trivial since it imposes a certain top-down structure on the data that can make it more or less difficult to extract insights depending on what features are used. Nonetheless, we found that the FINE embedding using all handpicked features contains rich information regarding how different clusters of communities were affected by and responded to (typically) climate-related shocks. This enabled us to identify key drivers to explain why certain communities were clustered together and to identify underlying human behavioural patterns. Applying FINE to other real-world datasets would be highly instructive regarding its capabilities and limitations. While not being limited by the need to impose a grouping scheme on data, t-SNE and Diffusion Maps require hyperparameter tuning and thus more computational effort, unlike FINE which is non-parametric. Another drawback is that t-SNE and Diffusion Maps do not seem to handle missing feature values well, which caused us to remove a significant portion of households in order to generate the embeddings displayed in Figure 6. FINE on the other hand can simply use all available values to estimate group probability mass functions. Nonetheless, t-SNE and Diffusion Maps allow clusters to emerge in a more bottom-up way, which can be desirable when a natural grouping of observations is not clear.

The choice of algorithm ultimately depends on the researcher's goals. t-SNE and Diffusion Maps may be more suitable for exploratory data analysis and for discovering whether data contains any intrinsic clusters. On the other hand, when a grouping scheme is obvious or supported by existing literature, then FINE seems to be a more suitable and straightforward choice. Of course, using a combination of these approaches is possible, and in fact can help to extract greater insight from data, as well as ensure that results are robust across different methods.

Acknowledgements The authors acknowledge the support from the Netherlands eScience project of the Dutch NWO, under contract 27020G08, titled "Computing societal dynamics of climate change adaptation in cities".

Data and Code Availability Data from the Bangladesh Climate Change Adaptation Survey is available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/27883. All code used in this paper can be found at: https://github.com/charlesaugdupont/cca-manifold-learning

Supporting Information Supporting tables and figures can be found at: https://github.com/charlesaugdupont/cca-manifold-learning/blob/main/SI.pdf

References

1. Carter, K.M., Raich, R., Finn, W.G., Hero III, A.O.: Fine: Fisher information nonparametric embedding. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(11), 2093–2098 (2009). https://doi.org/10.1109/TPAMI.2009.67
2. Cayton, L.: Algorithms for manifold learning. Univ. of California at San Diego Tech. Rep **12**(1-17), 1 (2005)
3. Cerda, P., Varoquaux, G., Kégl, B.: Similarity encoding for learning with dirty categorical variables. Machine Learning **107**(8), 1477–1494 (Sep 2018)
4. Coifman, R.R., Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis **21**(1), 5–30 (2006). https://doi.org/https://doi.org/10.1016/j.acha.2006.04.006
5. Delaporte, I., Maurel, M.: Adaptation to climate change in bangladesh. Climate policy **18**(1), 49–62 (2018)
6. Fodor, I.K.: A survey of dimension reduction techniques. Tech. rep., Lawrence Livermore National Lab., CA (US) (2002)
7. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (Mar 1975). https://doi.org/10.1007/BF02291478
8. Har-Shemesh, O., Quax, R., Lansing, J.S., Sloot, P.M.A.: Questionnaire data analysis using information geometry. Scientific Reports **10**(1), 8633 (Dec 2020). https://doi.org/10.1038/s41598-020-63760-8
9. Hinton, G., Roweis, S.: Stochastic neighbor embedding. Advances in neural information processing systems **15**, 833–840 (2003), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.7959&rep=rep1&type=pdf
10. (IFPRI), I.F.P.R.I.: Bangladesh Climate Change Adaptation Survey (BCCAS), Round II (2014). https://doi.org/10.7910/DVN/27883
11. Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**(1), 1–27 (1964)
12. Reckien, D., Creutzig, F., Fernandez, B., Lwasa, S., Tovar-Restrepo, M., Mcevoy, D., Satterthwaite, D.: Climate change, equity and the sustainable development goals: an urban perspective. Environment and Urbanization **29**(1), 159–182 (2017). https://doi.org/10.1177/0956247816677778
13. Scargle, J.D.: Studies in astronomical time series analysis. v. bayesian blocks, a new method to analyze structure in photon counting data. The Astrophysical Journal **504**(1), 405–418 (sep 1998). https://doi.org/10.1086/306064
14. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research **9**(nov), 2579–2605 (2008)