Manifold Analysis for High-Dimensional Socio-Environmental Surveys

Charles Dupont and Debraj Roy

Faculty of Science, Informatics Institute, University of Amsterdam,
Science Park 904, 1090, GH, 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. However, the above methods either rely on a pairwise correlation matrix, assume specific (conditional) probability distributions in its construction, or assume that the high-dimensional 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 two methods except for questionnaires with high multi-partite information. While not being limited by the need to impose a grouping scheme on data, t-SNE and Diffusion Maps require some tuning and thus more computational effort since they are sensitive to the choice of hyperparameters, unlike FINE which is non-parametric. Finally, we show that FINE is able 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) [11]. Current global warming trends and their adverse impacts such as floods represent a complex problem, which cannot be understood independently of their socioeconomic, political, and cultural contexts. Extreme weather events and increased climate variability have adversely impacted agriculture and food security, leading to global reductions in production and livelihood options. For example, communities dependent on agricultural land, local fisheries and other commons often have low incomes, which reinforces individual vulnerability and leads to poverty traps in the event of extreme climate events. The impact of climate change on farmers and their livelihoods is at a critical juncture and adaptation is proving to be 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 such high dimensional datasets, often in the form of questionnaires.

Although standard methods such as Principal Component Analysis, Factor Analysis or Regression models are commonly 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 [5]. 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 [10], Diffusion Maps [3], and Fisher Information Non Parametric Embedding (FINE) [1]. We start by extending prior work [6] 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 [8], 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 any behavioural regimes of adaptation can be detected, 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 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 although it allows for the detection of outliers and adaptation regimes, while t-SNE and Diffusion Maps allow clusters to emerge more naturally since no prior structure is assumed. 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

In this section we introduce the three non-linear manifold learning algorithms of interest as well as a simulation framework for generating questionnaires. We describe a real world questionnaire pertaining to climate change adaptation in Bangladesh and provide an overview of our experiments.

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 [10], and is based on prior work on Stochastic Neighbour Embedding [7]. Key steps are presented and summarized in Algorithm 1. First, a probability distribution over pairs of data points in the original feature space is constructed in order to capture some notion of similarity, where the similarity of some data point $\mathbf{x}_j$ to data point $\mathbf{x}_i$ is defined as

$$p_{j|i} = \frac{\exp\left(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2\right)}. \tag{1}$$

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 certain perplexity value specified by the user, and calculated as perplexity$(P_i) = 2^{H(P_i)}$, where $H(P_i)$ is the Shannon entropy [10]. Because typically $p_{j|i} \neq p_{i|j}$, we define the joint distribution

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

where N denotes the total number of observations in the dataset.

The next step is to construct another probability distribution over the data, but this time in a lower-dimensional space with the aim of minimizing the KL divergence between the previous probability distribution and this newly constructed one. We thus seek a lower-dimensional representation of our data such that the similarities between data points in the original space are preserved. The joint probabilities for data points in this lower dimensional map are

$$q_{ij} = \frac{\left(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2\right)^{-1}}{\sum_{k\neq l}\left(1 + ||\mathbf{y}_k - \mathbf{y}_l||^2\right)^{-1}}, \tag{3}$$

which is a Student t-distribution with a heavy tail [10]. The KL divergence is minimized iteratively by means of gradient descent where $\mathbf{y}_i$ are updated at each step.

```
Algorithm 1: t-SNE
  Data: $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$
  Input: dimension $d$, perplexity, learning rate $\eta$, number of steps $T$, momentum $\alpha(t)$
  Result: $Y^{(T)}$, a lower-dimensional representation of the data
  begin
      Compute $p_{j|i}$, $p_{i|j}$ using Equation (1) and specified perplexity;
      Compute $p_{ij}$ using Equation (2);
      Initialize sample solution $Y^{(0)} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$ with $\mathbf{y}_i \in \mathbb{R}^d$;
      for $t = 1, 2, \ldots, T$ do
          Compute $q_{ij}$ using Equation (3);
          Compute gradient $\frac{\partial C}{\partial Y} = \frac{\partial}{\partial Y} KL(P||Q)$;
          Update $Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial Y} + \alpha(t)\left(Y^{(t-1)} - Y^{(t-2)}\right)$;
      end
  end
```

Diffusion Maps Diffusion Maps is a method introduced by Coifman and Lafron which takes inspiration from the processes of heat diffusion and random walks [3]. 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 between pairs of data points using a Gaussian kernel as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\epsilon}\right), \tag{4}$$

where ϵ is some normalization parameter. Subsequently, a diffusion process is constructed, which allows us to map diffusion distances to a lower-dimensional space. Algorithm 2 summarizes the important steps of this process.

Algorithm 2: Diffusion Maps

```
Data: $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$
Input: dimension $d$, $\alpha$, $\epsilon$, number of iterations $t$
Result: $\Psi_t$, a lower-dimensional representation of the data
begin
    Compute diffusion matrix $L_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ using Equation (4);
    Construct diagonal matrix $D_{i,i} = \sum_j L_{i,j}$ from row sums of $L_{i,j}$;
    Normalize the matrix: $L^{(\alpha)} = D^{-\alpha} L D^{-\alpha}$;
    Compute $M = (D^{(\alpha)})^{-1} L^{(\alpha)}$, where $D^{(\alpha)}_{i,i} = \sum_j L^{(\alpha)}_{i,j}$;
    Compute $d$ largest eigenvalues $(\lambda_1^t, \ldots, \lambda_d^t)$ of $M^t$ and corresponding eigenvectors
     $(\psi_1, \ldots, \psi_d)$;
    Construct embedding $\Psi_t(\mathbf{x}_i) = (\lambda_1^t \psi_1(\mathbf{x}_i), \ldots, \lambda_d^t \psi_d(\mathbf{x}_i))$;
end
```

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. [6]. Algorithm 3 summarizes key steps. First, the authors divide respondents to the questionnaire 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 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 [9]. Note that questionnaire items are assumed to be independent such that probabilities may be factorized.

Algorithm 3: 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 $M_{ij} = \cos^{-1}\left(\sum_I \xi_I^{(j)} \xi_I^{(k)}\right) \to$ arc length on the hypersphere;
    end
    Construct embedding $E = \mathrm{MDS}(M, d)$;
end
```

2.2 Simulation Framework

Framework Description The authors of [6] 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

$$\phi_i^\kappa(t) = \begin{cases} \frac{\pi}{2}\sin^2(m\pi t), & i = \kappa \\ \frac{\pi}{2}t, & i \neq \kappa \end{cases}, \tag{5}$$

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:

$$\begin{aligned} \xi_1 &= \cos(\phi_1) \\ \xi_2 &= \sin(\phi_1)\cos(\phi_2) \\ \xi_3 &= \sin(\phi_1)\sin(\phi_2)\cos(\phi_3) \\ &\vdots \\ \xi_{N-1} &= \sin(\phi_1)\ldots\sin(\phi_{N-2})\cos(\phi_{N-1}) \\ \xi_N &= \sin(\phi_1)\ldots\sin(\phi_{N-2})\sin(\phi_{N-1}) \end{aligned} \tag{6}$$

For some choice K, which denotes the total number of groups (see Algorithm 3), we draw K values uniformly on the curve given by Equation (5). Then, we compute probabilities $p_I^{(k)} = (\xi_I^{(k)})^2$ and randomly generate a number of questionnaire responses for each group $k = 1, 2, \ldots, K$.

Experiments We wish to compare the embeddings generated by t-SNE, Diffusion Maps, and FINE for various simulated questionnaire responses. Similarly to [6], 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, similarly to what is done in [6], we apply the Procrustes algorithm, 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 (6).

Parameter Tuning FINE does not require any parameterization, although a grouping scheme must be provided. Such a scheme is explicitly defined in 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 $\alpha = \frac{1}{2}$.

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 [8]. This survey is a follow-up to a first round of data collection, which was studied in detail in [4]. The present work similarly focuses on household attributes such as expenditure, assets, adverse weather events, changes in farming practices, group membership and social capital. Some households have frequently missing response fields, so we retain 805 households having responded to at least 30% of survey questions.

Spatial Characteristics Localization codes corresponding to the district, "upazila", union, and village that each household belongs to are provided in the survey. There are 40 communities in total, each one being defined as a unique combination of the district, "upazila", and union codes. Each community contains an average of 20.125 households. Each household also possesses one of 7 distinct codes corresponding to different agro-ecological zones. Only agro-ecological and district codes are used for our analysis, although other localization codes are used to bin households into communities. Spatial characteristics are summarized in SI Table 2.

Adaptation Options Households were asked what changes they have implemented in response to climate change. We consider 31 different binary adaptation options, summarized in SI Table 3. Approximately 91% of households adapted in at least one regard, the most implemented adaptation being to change crop varieties (76% adoption rate).

Climatic Shocks Households were also asked whether various weather events adversely impacted the household's activities. There are 13 weather event categories, yielding 13 binary features. Three more specific questions pertaining to weather-related shocks are also included. A summary of these features is presented in SI Table 1. Households were most frequently affected by droughts (56%) and flashfloods (26%). Only 2 households reported being impacted by changing seasons, and just 1 by sea level rise.

Household Characteristics Various pieces of information were collected about household characteristics. Continuous features include household size, household head age, number of assets and value, number of lands and value, owned quantities of various livestock, as well as expenditure and income. SI Table 4 provides summary statistics of these features. Discrete binary features include: sex of household head (1 for female, 0 for male), whether the primary and secondary occupations of the household head are in agriculture (1 if so, 0 if not), and whether any member of the household has pursued religious education (1 if so, 0 if not). SI Table 5 provides a summary of these features. Lastly, categorical features include the highest education level amongst household members (SI Table 6), and the type of access to electricity, if any (SI Table 7).

Handpicked Features We also construct a set of handpicked features expected to be important for determining adaptation strategies based on existing literature. These features include: household income and expenditure, occupations of household members, losses due to weather and personal shocks, what actions were taken in response, social capital, collective action, constraints to adaptations and what adaptations were implemented, and finally what community groups household members are a part of as well as any associated benefits. This set of features is summarized in SI Table 8.

Mutual Information Based Feature Selection Finally, we also use a mutual-information based approach to feature selection by extracting features having high mutual information with target features associated with adaptation. The severity level of each constraint to adaptation (no access to credit, land, input, or money, scarcity of water or labor, no market, no information on climate change and appropriate adaptations, "other") is taken in turn as the target, each feature receives its average mutual information associated with the targets as overall score. The community that households are a part of is the highest-scoring feature, further motivating the choice of forming groups at the community level for FINE. In addition to the community, we retain the top 30 features. SI Table 9 summarizes this set of features.

Discretisation of Continuous Features We choose to discretise continuous features into bins in order to facilitate the estimation of probability mass functions. This is done using the dynamic programming method of bayesian blocks, first introduced by Scargle [12]. This approach seeks an optimal segmentation of a set of continuous data, yielding bins of different sizes, each of which corresponds to a separate uniform distribution. We choose to discretise continuous features into at most 5 bins. SI Figure 1 shows a comparison of a raw histogram and bayesian blocks for average monthly household income and expenditure.

Experiments We apply FINE to each set of feature described in this section. For handpicked features and mutual information based features, we apply FINE both for the community grouping, as well as a randomized grouping of households. We also 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

In this section we present the results obtained for the various experiments carried out on the simulation framework as well as the climate change adaptation questionnaire.

3.1 Simulation Framework Results

Figure 1 displays the theoretical embedding for each (κ, K) pair, along with the embeddings obtained using t-SNE, Diffusion Maps, and FINE for one sample questionnaire. t-SNE performs remarkably well for the $(\kappa = 1, K = 20)$ questionnaire, though its performance is lower than FINE and Diffusion Maps for the other (κ, K) pairs. FINE performs best out of all algorithms and the embeddings it produces thus also visually resemble the theoretical embeddings more closely.

The multi-partite information is also displayed, which is a measure for the amount of dependence between the questions of the simulated questionnaires, and is defined as

$$\mathrm{MI} \equiv \sum_I p_I(q_1, q_2, \ldots, q_{N_Q}) \ln \frac{p_I(q_1, q_2, \ldots, q_{N_Q})}{p_I(q_1)p_I(q_2)\ldots p_I(q_{N_Q})}. \tag{7}$$

Figure 2 displays the distribution of performance (correlation with theoretical embedding) of each algorithm for all 30 questionnaires and each (κ, K) combination. For each algorithm and (κ, K) combination we report only the results for the best-performing hyperparameters, with the exception of FINE which is non-parametric and only has one set of results. 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).

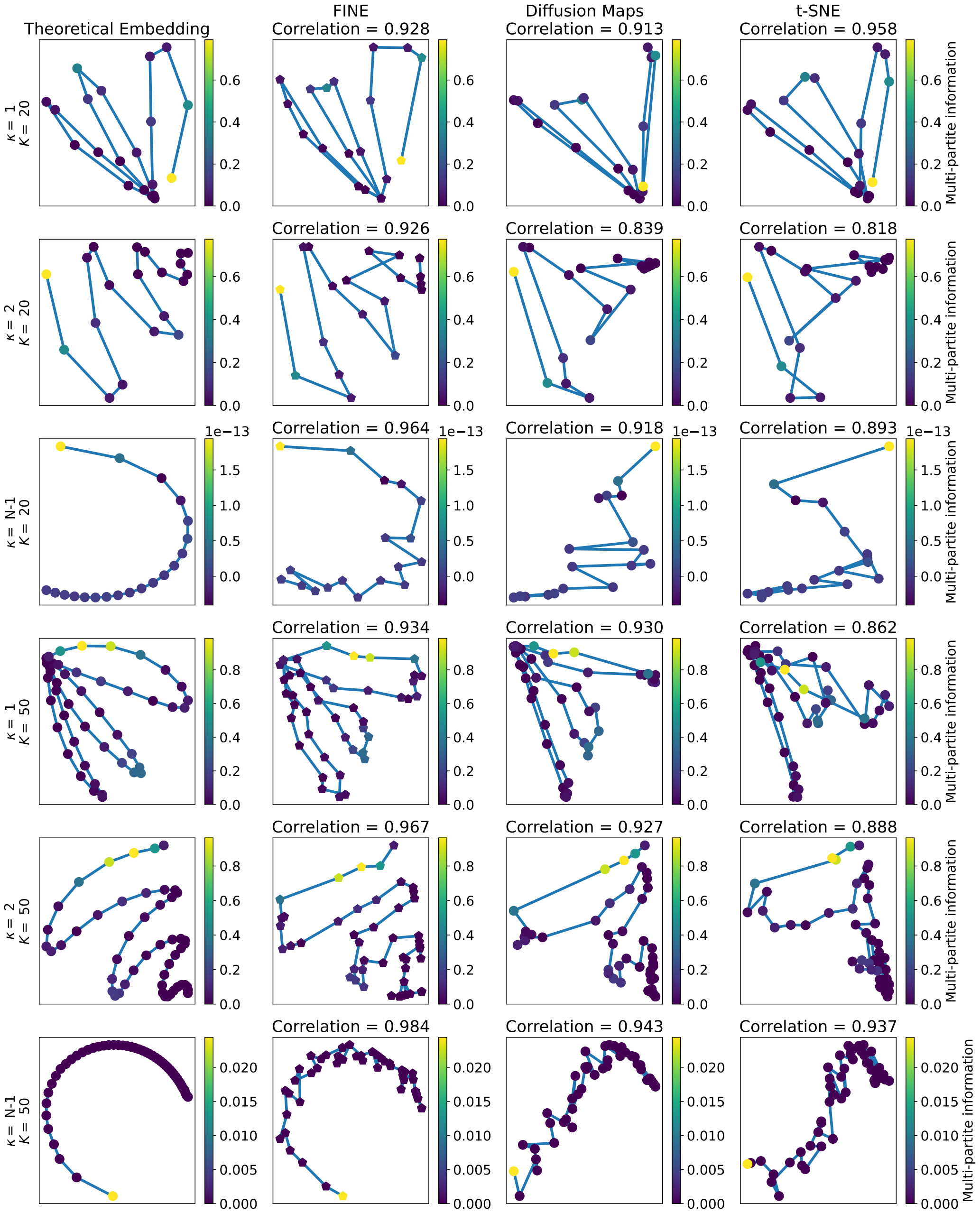

Fig. 1: Comparison of the embeddings obtained for t-SNE, Diffusion Maps, and FINE with respect to the theoretical embedding using the simulation framework. Each column corresponds to a different algorithm, and each row corresponds to a different (κ, K) pair. t-SNE achieves comparable performance to FINE due to hyperparameter tuning, which is a departure from results presented in [6].

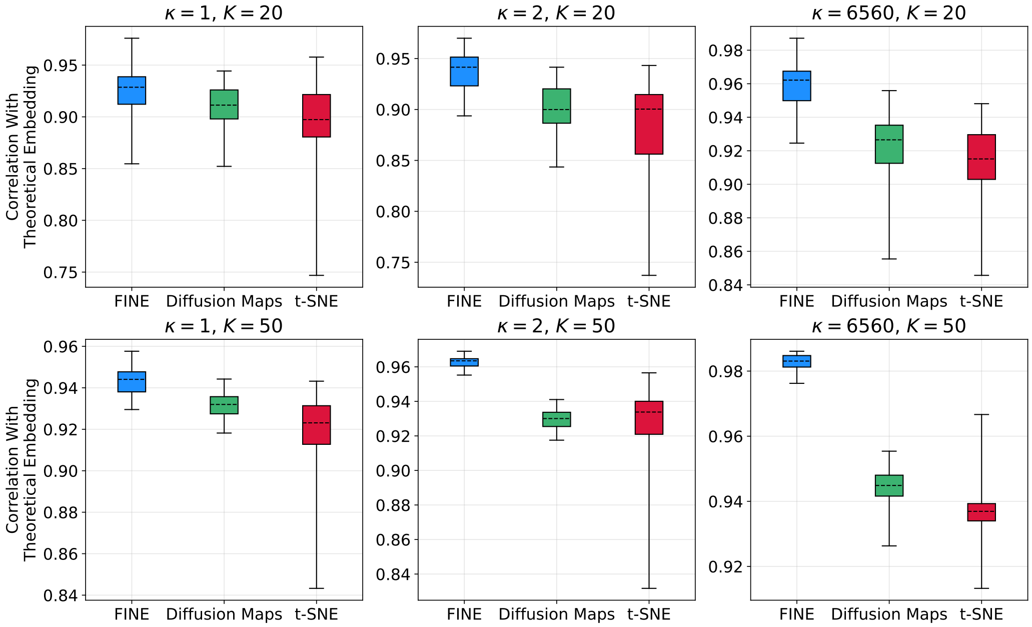

Fig. 2: Distributions of correlation coefficients with respect to the theoretical embedding for each algorithm and (κ, K) pair using best hyperparameters. Dashed lines denote mean values, and whiskers denote the maximum and minimum values.

Table 1 displays the best hyperparameter combinations for t-SNE. Lower perplexity values typically perform better, as does the weighted hamming distance. Not all hyperparameter combinations were found to yield good performance, emphasizing the importance of hyperparameter tuning for t-SNE.

Table 1: Summary of best hyperparameters for t-SNE.

	t-SNE Best Hyperparameters
$(\kappa = 1, K = 20)$	perplexity $= 2$, $\eta = 50$, weighted hamming (no one-hot)
$(\kappa = 1, K = 50)$	perplexity $= 5$, $\eta = 200$, weighted hamming (no one-hot)
$(\kappa = 2, K = 20)$	perplexity $= 2$, $\eta = 50$, weighted hamming (one-hot)
$(\kappa = 2, K = 50)$	perplexity $= 10$, $\eta = 200$, weighted hamming (one-hot)
$(\kappa = 6560, K = 20)$	perplexity $= 2$, $\eta = 50$, weighted hamming (one-hot)
$(\kappa = 6560, K = 50)$	perplexity $= 1$, $\eta = 50$, weighted hamming (one-hot)

Table 2 similarly displays the best hyperparameter combinations 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, we find that Diffusion Maps also requires some hyperparameter tuning in order to ensure that we achieve a strong performance.

Table 2: Summary of best hyperparameters for Diffusion Maps.

	Diffusion Maps Best Hyperparameters
$(\kappa = 1, K = 20)$	$\epsilon = 1.5$, $t = 0.5$, cosine (one-hot)
$(\kappa = 1, K = 50)$	$\epsilon = 1.5$, $t = 0.5$, cosine (one-hot)
$(\kappa = 2, K = 20)$	$\epsilon = 0.5$, $t = 0.5$, cosine (one-hot)
$(\kappa = 2, K = 50)$	$\epsilon = 0.5$, $t = 0.5$, cosine (one-hot)
$(\kappa = 6560, K = 20)$	$\epsilon = 2.0$, $t = 0.5$, cosine (one-hot)
$(\kappa = 6560, K = 50)$	$\epsilon = 2.0$, $t = 0.0$, cosine (one-hot)

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 $m \in \{1, 2, \dots, 10\}$. 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 2, we select $\epsilon = 0.5$ and $t = 0.5$, and use the cosine distance metric after one-hot encoding.

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 2, 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.

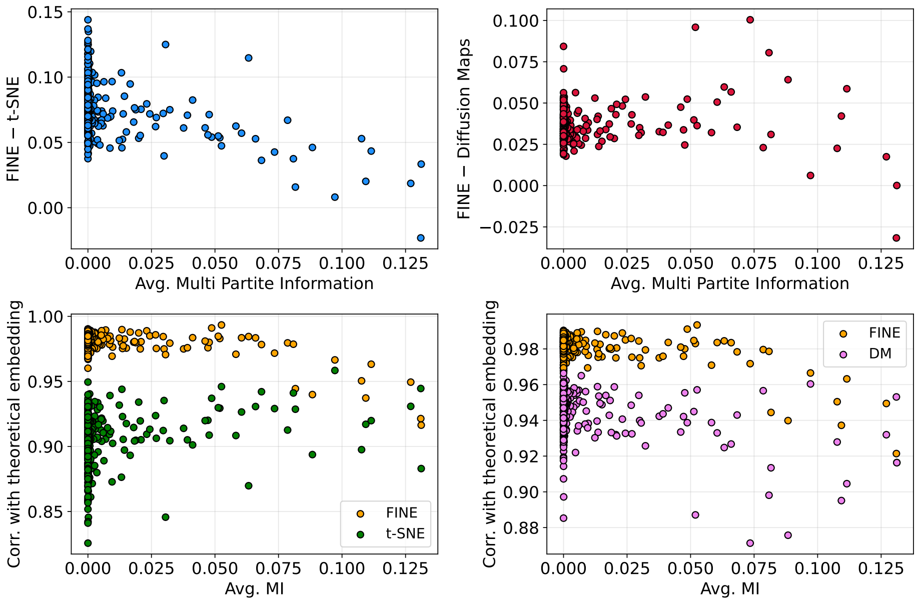

Fig. 3: (Top row) difference between FINE performance and t-SNE (left) as well as Diffusion Maps (right) as a function of multi-partite information, abbreviated MI. (Bottom row) FINE's performance begins to degrade for increasing multi-partite information. t-SNE (left) and Diffusion Map's (right) performances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

3.2 Bangladesh Climate Change Adaptation Survey Results

The FINE embeddings obtained for the various sets of features described in Section 2.3 can be viewed in SI Figure 2. Missing feature values were omitted from the estimation of probability mass functions. A few clusters emerge for most feature sets, the clearest of all being the embedding with spatial characteristics, where each agro-ecological code directly translates into a cluster of several communities. On the other hand, the randomized groupings for handpicked features and mutual information features produce embeddings with an almost uniform, isotropic distribution of groups, suggesting a sensitive dependence on the choice of grouping scheme. We shall focus mainly on the handpicked feature set (closely linked with adaptation) as an example, but note that the following analyses can be carried for other feature sets as well.

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 plot 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. Interestingly, community 21 also appears as a strong outlier. 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. The subsequent plots result in greater and greater dispersion of communities and less overlap, until we recover the "handpicked features" plot from SI Figure 2 once all 95 features have been added.

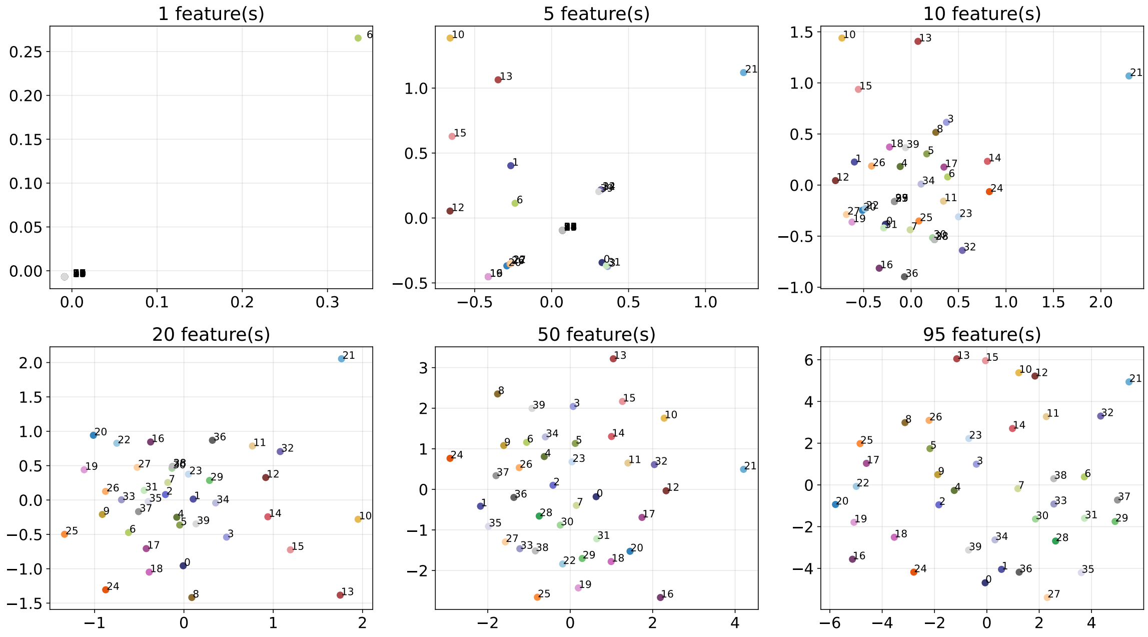

Fig. 4: FINE embeddings using an increasing number of handpicked features. Features are added in order from lowest to highest median KL divergence.

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 individual handpicked features. Some pairs of communities were excluded during the KL divergence computation, as were some features when approximating differential entropy due to numerical errors arising because of input values being nearly all zero. 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.

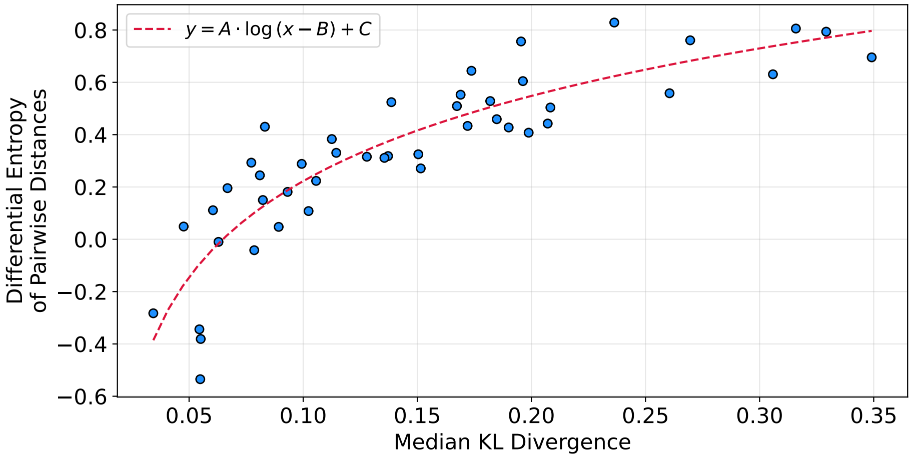

Fig. 5: Dependence of the (differential) entropy of pairwise distances in FINE embeddings on the median KL divergence between communities for handpicked feature. The dashed red line is a line of best fit with parameters: $A = 0.422, B = 0.014, C = 1.26$.

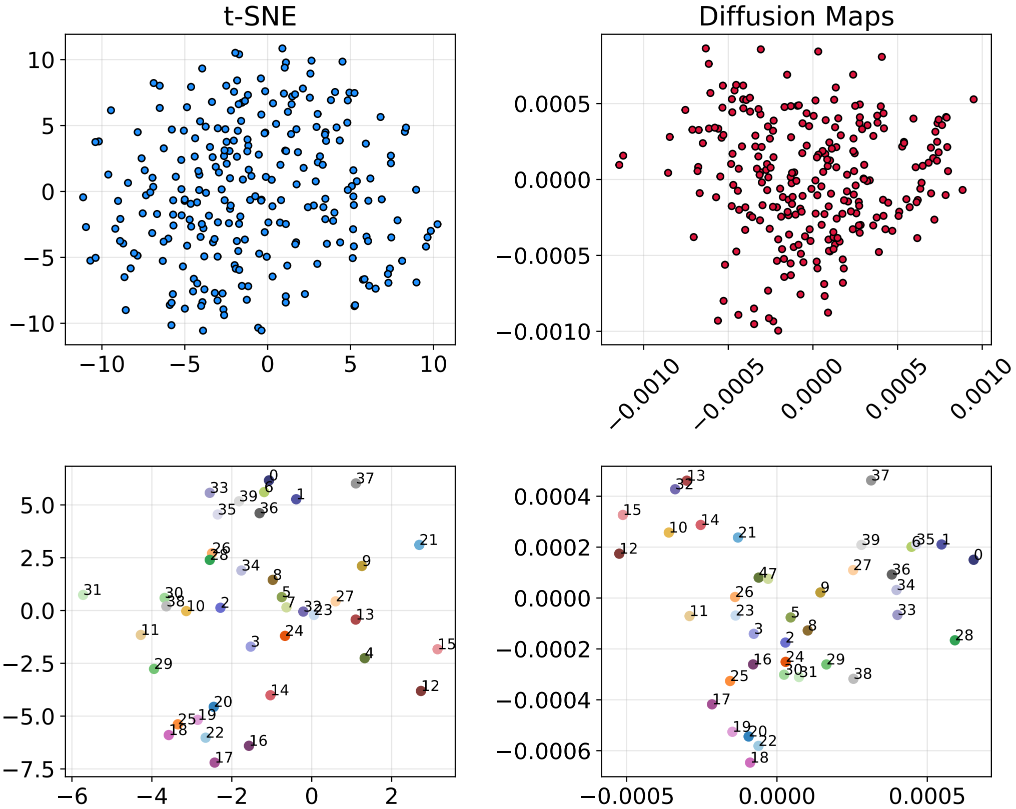

Fig. 6: (Top left) t-SNE embedding using handpicked features after one-hot encoding and using a weighted hamming distance metric. (Top right) Diffusion Maps embedding for same feature set after one-hot encoding and using the cosine distance metric. (Bottom row) community barycenters for t-SNE (left)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

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. In the bottom row, we plot the community barycenters by collapsing households belonging to the same community to their mean coordinates. Some interesting similarities can be found across the t-SNE and Diffusion Maps results. For example, communities 18, 19, 20, and 22 appear close together, as well as the pair 12 and 15. Note that these groupings also appear in Figure 4 in the bottom rightmost subplot.

As a final comparison with FINE, we plot pairwise distances between community coordinates for each pair of the three algorithms studied. All 95 handpicked features are used. Figure 7 shows these results. As we can see, overall there is agreement across all three methods regarding the topology and arrangement of the communities in relation to one another. Pearson correlation coefficients from left to right are: 0.569, 0.514, and 0.318. Perhaps unsurprisingly, correlation is highest between t-SNE and Diffusion Maps since the community grouping scheme was not applied to these two methods the way it was for FINE, and many households were omitted due to missing feature values.

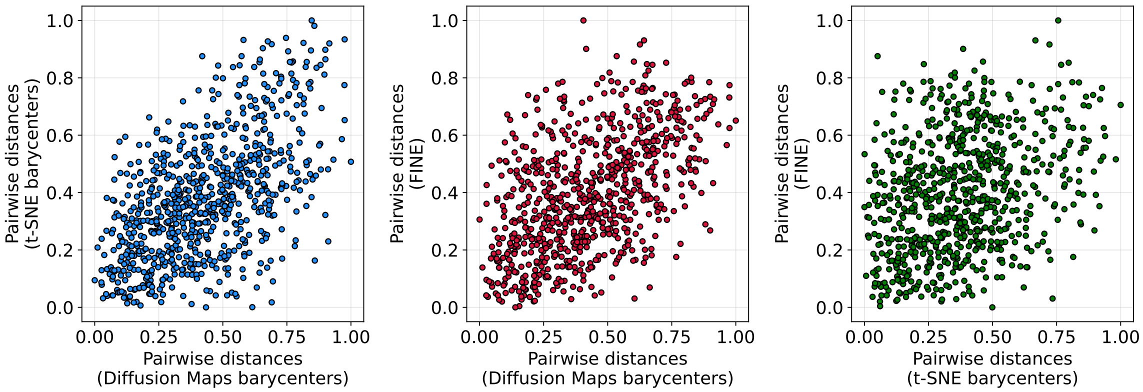

Fig. 7: Comparison of pairwise distances between community coordinates for each pair of algorithms using the handpicked feature set. Axes have been normalized between 0 and 1.

4 Discussion

In this section we provide a discussion and synthesis of key results as well as future lines of research. In the previous sections we have compared three non-linear manifold learning techniques for questionnaire data analysis, namely t-SNE, Diffusion Maps, and Fisher Information Non Parametric Embedding.

Experiments carried out with a simulation framework reveal that all three methods achieve comparable performance in terms of recovering the general shape of the underlying one-dimensional manifolds of interest. 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. Interestingly, FINE appears to consistently outperform the other two methods except for questionnaires with high multi-partite information. This reduction in performance is presumably due to the assumption that FINE makes about independence between survey items, which could potentially be relaxed for the more strongly-correlated survey questions. It remains to be seen exactly how FINE responds to a wider range of MI values. Pathways to simulating questionnaires with higher

MI values include more closely investigating the dependence of the MI values obtained on the various parameters of the framework, or the injection of dependencies by duplicating feature columns and introducing varying degrees of noise. Estimating multi-partite information for high-dimensional datasets, such as the climate change adaptation survey studied in this paper, is also challenging since the product of marginals in Equation (7) would quickly tend to zero.

The embeddings obtained with FINE reveal a logarithmic convergence in terms of how much dispersion of the groups can be observed in the embeddings as a function of the variability of the feature distributions across groups. 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 utilised. However, it can also allow for powerful detection of adaptation regimes. We saw earlier that communities 6 and 21 stood out in particular, and were able to identify key drivers by studying the embeddings generated for individual features. Imposing a grouping scheme therefore makes it possible to automatically detect divergent or unique regimes of behaviour for specific groups of observations.

While not being limited by the need to impose a grouping scheme on data, t-SNE and Diffusion Maps require some tuning and thus more computational effort since they are sensitive to the choice of hyperparameters 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 just use the values that are available for each household in order to contribute to the group probability mass function estimates. Nonetheless, t-SNE and Diffusion Maps allow for clusters to emerge in a more bottom-up way, which may 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 inherent or natural groupings. 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 algorithms.

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 (Round 2) 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, A.O.: FINE: Fisher Information Non-parametric Embedding (Feb 2008), http://arxiv.org/abs/0802.2050, arXiv:0802.2050 [stat]
2. Cayton, L.: Algorithms for manifold learning. Univ. of California at San Diego Tech. Rep **12**(1-17), 1 (2005)
3. 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, https://www.sciencedirect.com/science/article/pii/S1063520306000546, special Issue: Diffusion Maps and Wavelets
4. Delaporte, I., Maurel, M.: Adaptation to climate change in bangladesh. Climate policy **18**(1), 49–62 (2018)
5. Fodor, I.K.: A survey of dimension reduction techniques. Tech. rep., Lawrence Livermore National Lab., CA (US) (2002)
6. 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, http://www.nature.com/articles/s41598-020-63760-8
7. 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
8. (IFPRI), I.F.P.R.I.: Bangladesh Climate Change Adaptation Survey (BCCAS), Round II (2014). https://doi.org/10.7910/DVN/27883, https://doi.org/10.7910/DVN/27883
9. Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**(1), 1–27 (1964)
10. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008), http://www.jmlr.org/papers/v9/vandermaaten08a.html
11. 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, https://doi.org/10.1177/0956247816677778
12. Scargle, J.D.: Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data. **504**(1), 405–418 (Sep 1998). https://doi.org/10.1086/306064