Fair Neighbor Embedding

Jaakko Peltonen [*1] **Wen Xu** [*1] **Timo Nummenmaa** [1] **Jyrki Nummenmaa** [1]

Abstract

We consider fairness in dimensionality reduction (DR). Nonlinear DR yields low dimensional representations that let users visualize and explore high-dimensional data. However, traditional DR may yield biased visualizations overemphasizing relationships of societal phenomena to sensitive attributes or protected groups. We introduce a framework of fair neighbor embedding, the Fair Neighbor Retrieval Visualizer, formulating fair nonlinear DR as an information retrieval task with performance and fairness quantified by information retrieval criteria. The method optimizes low-dimensional embeddings that preserve high-dimensional data neighborhoods without biased association of such neighborhoods to protected groups. In experiments the method yields fair visualizations outperforming previous methods.

1. Introduction

Dimensionality reduction (DR) finds lower-dimensional representations for high-dimensional data sets, which can then be used as features for automated processing or can be explored by analysts. DR is applicable in numerous domains including high-dimensional data of society and individuals.

We consider unsupervised DR which aims to reveal structure of data without restricting to separation of pre-existing known classes. Many unsupervised DR methods have been introduced, from linear projections such as Principal Component Analysis (PCA: Pearson, 1901) to nonlinear DR approaches such as Stochastic Neighbor Embedding (SNE: Hinton & Roweis, 2002), t-distributed Stochastic Neighbor Embedding (t-SNE; Van der Maaten & Hinton, 2008), Neighbor Retrieval Visualizer (NeRV; Venna et al., 2010), Large Vis (Tang et al., 2016), and Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018). Some frameworks connecting such methods have been proposed, for example Assel et al. (2022) showed that several pairwise similarity DR methods can be stated in a common probabilistic model involving Markov Random Fields Graphs coupled by a cross entropy.

DR typically cannot preserve all original similarities or structures in high-dimensional data when the output dimensionality is lower than the effective dimensionality of data, as is usual in visualization, thus visualization methods must choose which data aspects to preserve. Typically they aim to preserve/reveal prominent variation or similarity structure in data. However, when run on data of human society this can yield visualizations that lack fairness as discussed next.

Data from societal phenomena, such as employment, education, banking, insurance, crime, are prominent in machine learning assisted analysis and decision making. However, variation (groupings, trends, similarities and other structures) in such data may seem related to sensitive attributes (protected groups) such as gender, race, age, religion and so on, even though dependencies may arise from biases in data collection or discrimination in the phenomena. To avoid discrimination in decision making, ethical artificial intelligence including fairness, accountability and transparency has become central in machine learning. Many works on fairness deal with automated decision making. We stress that fairness should extend to DR and visualization, so human analysts are not erroneously led to unfair conclusions while exploring data.

On societal data, traditional DR does not suffice for fairness. Methods like t-SNE can yield biased results showing "low-hanging fruit" of variation depending on sensitive attributes. Revealing it can be a first step of analysis, but traditional DR does not help explore beyond it: traditional DR does not reveal how variation could be explained by attributes other than sensitive ones, and may omit variation not explained by sensitive attributes. Thus analysts should complement traditional DR with visualizations by new, fair DR methods. *We define fair DR as methods that reveal whether variation can be explained without dependence on sensitive attributes.* We provide a first neighbor embedding solution for fair DR.

Naive solutions like leaving out sensitive attributes from DR input features do not suffice as fair DR: other variables may have statistical dependencies with sensitive ones, and remov-

[*]Equal contribution [1]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland. Correspondence to: Jaakko Peltonen <jaakko.peltonen@tuni.fi>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

ing all variables having dependency with sensitive attributes would reduce ability to visualize structure of the original data. New solutions are needed for low-dimensional embeddings that remove bias with respect to sensitive information, while preserving other structure of high-dimensional data.

We propose a framework called Fair Neighbor Retrieval Visualizer (Fair-NeRV), a generalization of NeRV that formulates fair dimensionality reduction as an information retrieval task. It preserves high-dimensional neighborhoods while avoiding dependency to sensitive attributes.

Illustrative example of fair DR. Consider synthetic data with 500 points along five dimensions, arising out of mixtures of Gaussians, and having a sensitive attribute with three categories (e.g. three ethnic groups). In dimensions 1-3 feature values arise from a three-component mixture with the component of each point determined by its sensitive category; in dimensions 4-5 feature values arise from another three-component mixture sampled independently of the sensitive attribute. Such data has $3 \times 3 = 9$ high-dimensional Gaussian clusters. Clustering along dimensions 1-3 is highly related to the sensitive attribute: showing it in DR output would reveal sensitive attributes of data points. In contrast, clusters along dimensions 4-5 can be preserved in DR without revealing sensitive information. A fair visualization should show only the latter clustering, not showing any clusters/subclusters related to sensitive categories.

Figure 1 shows plots by previous methods t-SNE and ct-SNE (Kang et al., 2021) and proposed methods Fair-NeRV and Fair-t-NeRV (columns a, b, c, d) applied to such data to produce two-dimensional outputs. The top row shows a coloring of points by their sensitive category: any cluster structure corresponding to the coloring indicates the data arrangement has revealed sensitive information, yielding non-fair visualization. The bottom row shows coloring of points by the high-dimensional mixture components of original dimensions 4-5: the better these colors are shown as clusters, the better the structure of the high-dimensional data is shown. A traditional t-SNE embedding (subfigures a and a') shows all 3×3 clusters, showing undesired association of data clustering to sensitive attributes. A recent variant ct-SNE (b and b') still reveals sensitive categories as color separation is visible in b. The two variants of our method (c and c', and d and d') successfully avoid showing any relationship to sensitive attributes (c and d), while fully revealing the remaining original high-dimensional structure (c' and d'; clusters correspond to mixture components of dimensions 4-5). This example uses cluster structure for ease of illustration, but the principle applies to any structure: fair visualization should show the high-dimensional structure as well as possible without showing dependencies to sensitive attributes. Such fair DR then complements traditional DR, to give analysts a comprehensive view of what biases exist in data and what structure can be explained without biases.

Our contributions: 1) We propose Fair-NeRV, a new nonlinear DR method preserving neighborhoods in data while removing dependency on sensitive information, thus preserving the remaining non-sensitive neighborhood structure. It includes fair versions of SNE and t-SNE as special cases. It involves a new fairness cost which can also be applied to other nonlinear DR methods. Ours is the first neighbor embedding solution designed for fair DR. 2) We propose an evaluation measure for fair DR performance. 3) In experiments our method outperforms comparable other methods.

2. Previous Work

We focus on **neighbor embedding methods for DR**, that optimize low-dimensional coordinates for data so that their neighborhood relationships approximate high-dimensional data neighborhoods. SNE defines conditional distributions of neighbors of a data point and measures differences by Kullback-Leibler divergences; t-SNE uses joint distributions of data and neighbors and Student-t based low-dimensional neighborhood probabilities.

The Neighbor Retrieval Visualizer (Venna et al., 2010) is a nonlinear DR method that aims to create low-dimensional data embeddings to preserve high-dimensional neighbor relationships of samples. It can be seen as a generalization of SNE/t-SNE, and relates its objective to information retrieval. Given the input data set $\{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^m$, the probability that data point j is picked as a neighbor of point i is

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

where σ_i^2 controls falloff of the p_{ij} with respect to distance.

NeRV outputs an embedding of the points $\{\mathbf{y}_i\}_{i=1}^n$, $\mathbf{y}_i \in \mathbb{R}^d$, in a d-dimensional output space (e.g., 2D/3D for visualization), where neighbor probabilities $q_i = \{q_{ij}\}_{j=1,\ldots,N,\ j\neq i}$ are defined based on output coordinates $\mathbf{y}_i$ so that the probability to pick j as a neighbor of i is

$$q_{ij} = \frac{\exp(-||\mathbf{y}_i - \mathbf{y}_j||^2/\sigma_i^2)}{\sum_{k\neq i} \exp(-||\mathbf{y}_i - \mathbf{y}_k||^2/\sigma_i^2)} \,. \tag{2}$$

NeRV optimizes the low-dimensional coordinates to minimize the difference between the high and low dimensional neighborhood distributions, as measured by Kullback-Leibler divergences. The objective function of NeRV is:

$$C_{\text{NeRV}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\lambda D_{KL}(p_i, q_i) + (1-\lambda) D_{KL}(q_i, p_i)\Big)$$

where λ is a tradeoff parameter, and D_{KL} is the Kullback-Leibler divergence, $D_{KL}(p_i, q_i) = \sum_{j\neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$. Divergences of neighborhoods in NeRV are evaluated in both

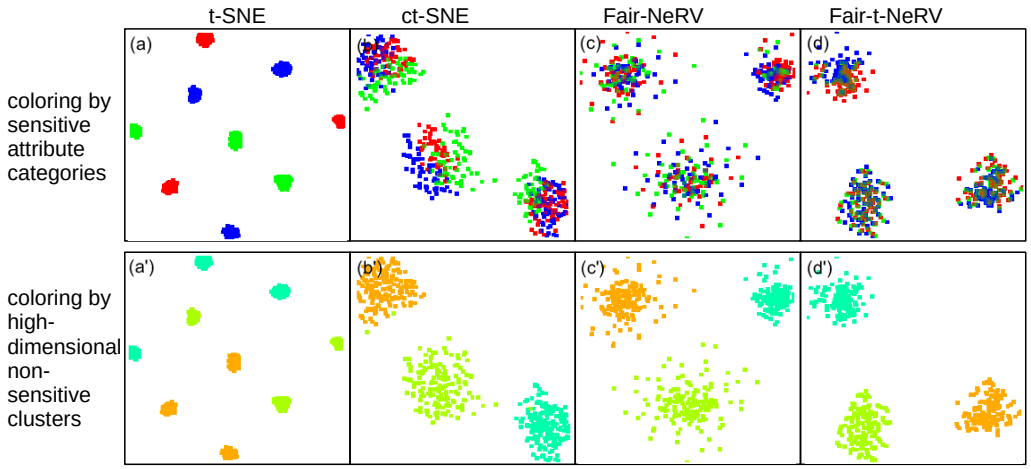

Figure 1. Dimensionality reduction results for SyntheticData. (a) and (a'): t-SNE result. (b) and (b'): ct-SNE result. (c) and (c'): Fair-NeRV result. (d) and (d'): Fair-t-NeRV result. In each column, colors in the top figure show the categorical values of the sensitive attributes, and colors in the bottom figure show a clustering of the original high-dimensional data (here a clustering independent of the sensitive attribute) as examples of its high-dimensional structure. A good visualization preserves the high-dimensional data structure, showing its clusters clearly, while not revealing any clustering of the sensitive attributes.

directions, from high-dimensional to low-dimensional space and vice versa. The objective has an interpretation as performance of an information retrieval task: the divergence $D_{KL}(p_i, q_i)$ has been shown to generalize an information retrieval cost of misses in retrieving original high-dimensional neighbors from the low-dimensional space, and the divergence $D_{KL}(q_i, p_i)$ is a generalization of the cost of false neighbors in the retrieval. The weight λ is a tradeoff of the costs of misses versus false neighbors. Thus NeRV optimizes embeddings for information retrieval of neighbors, minimizing the cost of errors. SNE/t-SNE can be seen as special cases of NeRV which minimize misses only.

NeRV has been extended to settings having class labels. The extension ClassNeRV (Colange et al., 2020) uses a cost function that separates neighbor distributions to within-class and between-class neighbors and penalizes their input-output space probability differences separately:

$$C_{\text{CNeRV}} = \sum_i \tau^{\in} D_B(p_i^{\in}, q_i^{\in}) + (1-\tau^{\in}) D_B(q_i^{\in}, p_i^{\in}) + \tau^{\notin} D_B(p_i^{\notin}, q_i^{\notin}) + (1-\tau^{\notin}) D_B(q_i^{\notin}, p_i^{\notin})$$

where $D_B(p_i^{S_i}, q_i^{S_i}) = \sum_{j \in S_i} p_{ij} \log \frac{p_{ij}}{q_{ij}} + q_{ij} - p_{ij}$ is a Bregman divergence between subsets of probabilities p_i and q_i and S_i is a set of indices summed over; in a Bregman divergence the probability subsets do not need to sum to 1. Here $S_i^{\in}$ is the set of within-class neighbors of i and $p_i^{\in} = \{p_{ij}\}_{j \in S_i^{\in}}$ and $q_i^{\in} = \{q_{ij}\}_{j \in S_i^{\in}}$ are the corresponding probability subsets, similarly $S_i^{\notin}$ are the between-class neighbors of i and $p_i^{\notin}$ and $q_i^{\notin}$ their probability subsets, and $(\tau^{\in}, \tau^{\notin}) \in [0,1]$ are tradeoff parameters for the two divergences in each subset. If $\tau^{\in} > \tau^{\notin}$, the ClassNeRV cost function penalizes more within-class missed neighbors and between-class false neighbors than other distortions; this was the setting in (Colange et al., 2020).

Previous work on fair DR. There have been recent works on fairness in DR. Fair PCA (Samadi et al., 2018) imposed constraints on learning a subspace for two protected groups by minimizing maximum deviation of reconstruction error for all protected groups. Olfat and Aswani (2019) proposed a quantitative definition of fairness for DR and developed a convex SDP formulation for Fair PCA. Later Kamani et al. (2022) introduced a Pareto Fair PCA and an algorithm to learn a subspace preserving fairness while slightly compromising reconstruction error. Methods such as MMD-based fair PCA (MBF-PCA; Lee et al., 2022), Iterative Null-space Projection (Ravfogel et al., 2020) and Relaxed Linear Adversarial Concept Erasure (Ravfogel et al., 2022) have been proposed. However, the above perform only linear DR. Adapting linear fair DR methods to nonlinear DR is challenging. Linear methods may rely on "null spaces", or projections defined by orthogonality notions that can remove the effect of a sensitive attribute entirely; in nonlinear mappings null-space mappings may not exist, let alone have an easily definable mathematical form. Thus nonlinear DR needs a new approach.

Demographic parity (Agarwal et al., 2018) and equal opportunity (Hardt et al., 2016) are common fairness notions for pairs of discrete-valued variables, i.e. target classes of a classifier and sensitive-attribute categories in a supervised setting. Although some linear methods above have connections to these measures, in general, in our exploratory data analysis setting output embeddings are continuous-valued

organizations of data (multivariate coordinates) that do not follow a simple standard distribution. Thus it is not feasible to compute demographic parity/equal opportunity directly for our setting and adapting such measures to the setting is nontrivial. However, the fairness cost we will introduce can be seen as a novel local demographic parity measure, as detailed in Appendix M.

Conditional t-SNE. There has not been much research in fair nonlinear DR, and in particular not in fair neighbor embedding. However, we now point out that a recent algorithm which was not introduced for the purpose of fair DR turns out to be usable for it. Conditional t-SNE (ct-SNE; Kang et al., 2021) is an extension of t-SNE that aims to find "complementary" embeddings where already-known structure is "factored out". It constructs a conditional low-dimensional neighborhood probability given class labels. The aim is to modify low-dimensional neighbor relationships to reduce importance of keeping same-labeled data points nearby.

We introduce a new way to use ct-SNE: if sensitive attribute categories are given as classes, it turns out the ct-SNE objective reduces the degree to which the method aims to keep points in the same sensitive category near each other in DR output. Thus ct-SNE can be used as a baseline for nonlinear fair DR. This is a novel use of ct-SNE. However, ct-SNE has downsides: its objective is mathematically opaque due to its modifications of low-dimensional neighbor probability, which no longer correspond to straightforward closeness on a display, and the ct-SNE objective has no clear information retrieval interpretation. We use ct-SNE as a baseline in experiments. Next we introduce our new nonlinear fair DR method which directly formulates fair DR as an information retrieval task, and outperforms ct-SNE in experiments.

3. Fair DR: Our Method

A simplistic approach to achieve fairness in DR might be removing the sensitive attributes from the feature set before training the DR model. However, this is insufficient, as it can leave data variation along other features having strong dependencies with the sensitive attributes. As an extreme case, if some variable is an identical copy of a sensitive attribute, leaving such variation in the data would allow easy identification of sensitive attribute values based on data positions in the DR output. More generally, any data variation visible after DR having statistical dependencies with sensitive attributes would lead to unfairness.

Data variation having dependencies with sensitive attributes may originate from any (or all) original features and may occur in any part of DR output. Thus, leaving out original features that are, e.g., correlated with sensitive attributes is not suitable: one might end up leaving out all features, or too many, so that little information remains to be visualized. Instead, DR should aim to create a mapping where visible variation in output is not associated with sensitive attributes. The mapping would need to be learned from data with knowledge of their sensitive attributes, so that their dependency with DR output can be evaluated and removed.

3.1. Fair-NeRV and Fair-t-NeRV

We develop our novel method by several steps. We first develop a novel fairness objective, and combine it with a novel use of the ClassNeRV neighbor retrieval objective.

We create an information retrieval based fairness objective, which can then be combined with a neighborhood preservation objective. The idea is: *if the visualization is fair, the position of a data point in the visualization should not reveal its sensitive attribute*. Thus, in a fair visualization one should not be able to predict the sensitive attribute value of a point based on its location. Therefore, we will design a cost function penalizing success of such prediction.

Neighbor-based prediction of sensitive attributes. For a neighbor embedding DR method it is appropriate to use neighbor based prediction of sensitive attributes. We consider a leave-one-out scheme, to predict the sensitive-attribute value of each data point i from its neighbors.

Let S be the number of sensitive attribute values. Consider a point i as a central point, meaning a point whose neighbors we aim to retrieve. Suppose we know the sensitive-attribute values for neighbors j but not the central point i. Based on neighbors of point i in the output space, the distribution $r_i = \{r_{is}\}_{s=0,\dots,S-1}$ can be estimated, where r_{is} is the estimated probability of sensitive-attribute value s. The probability can be estimated as

$$r_{is} = \frac{\sum_{j\neq i} \delta(s_j, s) \exp(-||\mathbf{y}_i - \mathbf{y}_j||^2/\sigma_i^2)}{\sum_{j\neq i} \exp(-||\mathbf{y}_i - \mathbf{y}_j||^2/\sigma_i^2)} \qquad (3)$$

where s_j denotes the sensitive-attribute value of point j, and $\delta(s_j, s) = 1$ if $s_j = s$ and zero otherwise.

Note that a simple classification penalty approach is not sufficient for a fairness objective: minimizing probability r_{is_i} of the correct categorical sensitive attribute value would aim to surround each point of some sensitive category by members of different categories, ignoring their relative proportions, and disallowing any neighbors of the correct category. Instead, we construct a fairness objective that aims to keep the distribution r_{is} for each point i close to a desired distribution.

Desired sensitive attribute distribution. To have a fair visualization, each data point i should have neighbors with sensitive values other than its own value s_i. Moreover, among the other sensitive values no data point should prefer a particular value among its neighbors. The desired distribution then has two properties: a) to control desired fairness,

the distribution should allow us to control the proportion of the categorical value s_i; b) the distribution should otherwise be uninformative about any other sensitive values, in other words it should follow overall proportions of those values.

The desired distribution for data point i is the same as the overall distribution of sensitive attribute values, except that we de-emphasize the prevalence of the sensitive value s_i of point i. We de-emphasize s_i since point i itself already brings that value into the neighborhood. Let $u = \{u(s)\}_{s=0,\dots,S-1}$ be the overall sensitive attribute distribution in the data set, computed by dividing counts of each sensitive attribute value by the number of data. We define a desired distribution $\rho_i = \{\rho_{is}\}_{s=0,\dots,S-1}$ of sensitive attribute values for each data point i as

$$\rho_{is} = \begin{cases} 1-\omega & \text{if } s = s_i \\ u(s)\cdot\omega/(1-u(s_i)) & \text{otherwise} \end{cases}$$

where $u(s_i)$ is the overall proportion of sensitive value s_i in the data and $\omega \in [0,1]$ is a weight controlling influence of value s_i in the neighborhood. When $\omega = 1-u(s_i)$ then ρ_i reduces to the overall sensitive value distribution, otherwise the larger the ω the more the prevalence of the value s_i is de-emphasized among the neighbors. For all choices of ω relative proportions of other sensitive values follow the overall distribution, as desired.

Information retrieval based fairness criterion. We define a fairness criterion measuring difference of the estimated sensitive attribute distribution around each point from their desired distributions. We measure this as an average of Kullback-Leibler divergences, each having an information retrieval interpretation: we set

$$C_{Fairness} = \frac{1}{N}\sum_i \Big(\gamma D_{KL}(\rho_i, r_i) + (1-\gamma) D_{KL}(r_i, \rho_i)\Big), \tag{4}$$

where for point i the divergence $D_{KL}(\rho_i, r_i)$ is a cost of *missed* sensitive attribute values (having smaller estimated probability than the desired one), $D_{KL}(r_i, \rho_i)$ is a cost of *false neighbor* sensitive values (having larger estimated probability than desired), and γ is a tradeoff parameter of the costs. Eq. (4) has an information retrieval interpretation:

Theorem 3.1. The cost $C_{Fairness}$ in (4) is an average information retrieval cost of misses and false neighbors. In detail, under simplifying assumptions, $C_{Fairness} \approx$

$$\frac{const.}{N}\sum_i \big(\gamma(1-recall(i)) + (1-\gamma)(1-precision(i))\big)$$

where $recall$ and $precision$ are the typical information retrieval measures, in the task of retrieving the desired distribution of sensitive attribute values for each point from the low-dimensional display. **Proof.** See Appendix B. This shows our fairness measure is directly task-based.

Fairness-aware neighbor retrieval criterion. It is natural to combine the information retrieval based fairness criterion $C_{Fairness}$ with the information retrieval based neighborhood preservation criterion of NeRV. However, they would be partly at odds: neighbor preservation would aim to preserve all neighbors of a point i, even those that would reveal its sensitive attribute. One should focus on preserving those neighborhood relationships that do not conflict with the fairness objective: we do this by modifying the ClassNeRV objective C_{CNeRV}. In ClassNeRV tradeoff parameters $(\tau^{\in}, \tau^{\notin})$ in C_{CNeRV} were set as $\tau^{\in} > \tau^{\notin}$ to penalize within-class missed neighbors and between-class false neighbors more than other distortions. In fair DR we want the *opposite effect*: we want to penalize missed neighbors of the same categorical sensitive value and false neighbors of other sensitive values less than other distortions, so that a fair DR output can arrange an equal mix of sensitive categories around each data point. Thus we can use C_{CNeRV} in an opposite fashion to achieve this, setting $\tau^{\in} < \tau^{\notin}$.

Final objective. To obtain fair visualization, we use a weighted combination of the ClassNeRV neighbor retrieval cost with $\tau^{\in} < \tau^{\notin}$, and the retrieval based fairness cost:

$$C_{FairNeRV} = \beta C_{CNeRV} + (1-\beta) C_{Fairness} \tag{5}$$

where β is a tradeoff parameter. The cost thus has an information retrieval interpretation as a sum of retrieval costs, with tradeoff parameters for which errors to penalize most.

Gaussian and t-distributed version. The $C_{FairNeRV}$ cost can be used with Gaussian low-dimensional distributions, or with Student-t based distributions as in t-SNE, where in eqs. (2) and (3) each exponent term $\exp(-a)$ is replaced by the form $1/(1+a^2)$. In the t-distributed version we also leave out divisions by σ_i^2 from (2) and (3) for similarity with methods such as t-SNE. We call the Gaussian version Fair-NeRV and the t-distributed version Fair-t-NeRV.

Special case. Setting $\tau^{\in} = \tau^{\notin} = 1$ yields a fair version of SNE and t-SNE as a special case. (In general our method optimizes $\tau^{\in}, \tau^{\notin}$.) The special case has similar performance to the general method in our experiments (Appendix I).

We note that the fairness cost can be combined with other dimension reduction methods such as Isomap or LLE by adding the cost to their objectives.

3.2. Optimization

The objective (5) is non-convex w.r.t the embedding Y; we find gradient based optimization from random initializations works well. Any gradient based optimizer can be used. Gradients have intuitive forms as forces between pairs of data; Gaussian version shown in Appendix A, the t-distributed one is similar. As in other pairwise methods, computational complexity is $\mathcal{O}(n^2)$ and can be made $\mathcal{O}(n \log n)$

with Barnes-Hut approximation; see Appendix C.

4. Experiments

We perform DR on six data sets, reducing them to 2D. We compare Fair-NeRV/Fair-t-NeRV with other methods and show output plots. To compare performance, we introduce a new goodness measure for fair DR. We split each data set to a training and test data set; parameters are optimized for training sets, performance is evaluated on test sets.

4.1. Comparison Methods

We compare our method Fair-NeRV/Fair-t-NeRV with nonlinear DR methods: t-SNE as a widespread method and conditional t-SNE (ct-SNE) since we showed its objective can be applied for fair DR. We give a comparison showing that Fair-NeRV/Fair-t-NeRV outperforms the linear fair DR method MBF-PCA in Appendix K, and comparisons to two restricted variants of our method (ablation studies) in Appendices I and J.

4.2. Data Sets

We use one synthetic data and five real data sets.

The *Syn* data set is an artificial set of 1000 data points with 5 dimensions. The first three dimensions have three multivariate Gaussian clusters, cluster membership considered sensitive information. Dimensions 4 and 5 dimension have an independent mixture of three multi-variate Gaussian clusters, considered non-sensitive information.
The *Adult* data (https://tinyurl.com/rsbk8wab), also called Census Income, is a fairness benchmark with 48842 instances and 14 attributes; gender is the sensitive attribute.
Communities and Crimes (CC; https://tinyurl.com/34sttp2t) is a socio-economic data set from 46 US states in 1990, with 1994 samples (communities) and 127 attributes (4 categorical and 123 numerical). The sensitive attribute is whether a community has over 6% African American population.
The *German credit* data (https://tinyurl.com/y6h95pne) has 1000 bank account holders with 13 categorical, 7 numerical and 1 binary attributes. Age is the sensitive attribute after binarization into {young, old} by age thresholding at 25.
The *Law School* (LSAC) data set stems from an admission council survey at 163 US law schools in 1991. It contains law school admission records of 20798 students with 12 attributes (3 categorical, 3 binary, 6 numerical); Race is considered as the sensitive attribute. We used preprocessed data from https://tinyurl.com/mtu84w83.
The *Pima* data (https://tinyurl.com/4ytz6nm8) has 8 medical predictor variables and a diabetes class variable for 768 patients. A BMI categorization is the sensitive attribute ($BMI < 25$: 0; $25 \leq BMI < 30$: 1; $BMI \geq 30$: 2).

For any data sets having over 1000 points we take 1000 points by stratified sampling with respect to sensitive categories. We then split each data set into 50% training and 50% test data again by stratified sampling. Sensitive attributes are excluded from input features for all methods.

4.3. Goodness Measures

While fairness measures have been proposed in tasks such as classification, suitable measures are needed for fairness and performance in nonlinear DR. We propose a measure.

Motivation: fairness depends on inspection scale. DR output can be inspected at different scales from local towards global, showing different levels of detail; DR performance of neighbor embedding methods is also often evaluated at different scales. We point out that *fairness in DR also depends on the inspection scale*: in e.g. k-nearest neighbor (kNN) based prediction of sensitive attribute values, as the neighborhood size k grows the distribution of sensitive values around any point becomes close to their overall distribution, and prediction performance falls towards that of random guessing (we discuss the measure we use in the next paragraph). Thus any DR output looks "fair" at a large enough scale as sensitive attributes cannot be predicted anymore. However, in an unfair visualization, data organization revealing sensitive attributes may be present at multiple scales, requiring a large k to reach random guessing.

Definition: Fair inspection scales. Let $\text{Perf}(k)$ be a performance measure of a scale-dependent sensitive attribute predictor, evaluated at scale k. We call k a *fair inspection scale* if prediction performance of sensitive attributes is at most a small fraction ϵ different from random guessing, i.e., $|\text{Perf}(k) - \text{Perf}(\infty)| \leq \epsilon$ where $\text{Perf}(\infty)$ is the limiting value corresponding to random guessing. We then define the *fairness scale threshold* k_{fair} as the smallest k such that all $k' \geq k$ are fair inspection scales:

$$k_{fair} = \min_k \text{ s.t. } \max_{k' \geq k} |\text{Perf}(k') - \text{Perf}(\infty)| \leq \epsilon \ .$$

Classification goodness measure. As sensitive attributes often have minority categories, hard kNN may have noisy behavior; instead we consider soft kNN classification, which assigns a point i to each sensitive value category $(s = j)$ with a weight $\epsilon_{i,j}$ corresponding to proportion of the k neighbors in the category. Denote the set of points having sensitive value j by Z_j. With a soft classification, precision of category j is defined as the proportion of all weight assigned to category j arising out of true members of the category, i.e., $\text{precision}(j) = (\sum_{i \in Z_j} \epsilon_{i,j})/(\sum_{i'} \epsilon_{i',j})$. Similarly, recall of category j is the proportion of weight at which the true members of the category are assigned to it, i.e., $\text{recall}(j) = (\sum_{i \in Z_j} \epsilon_{i,j})/|Z_j|$. The definitions reduce to usual precision and recall when the kNN classification becomes non-soft. From precision and recall the f1 score is

Table 1. Dimensionality reduction performances on test data sets by high dimensional cluster-based $f1_k$ and $f1_{\text{Avg}}$ scores.

Data	t-SNE			ct-SNE			Fair-NeRV			Fair-t-NeRV		
	k	$f1_k$	$f1_{\text{Avg}}$	k	$f1_k$	$f1_{\text{Avg}}$	k	$f1_k$	$f1_{\text{Avg}}$	k	$f1_k$	$f1_{\text{Avg}}$
Syn	388	0.3895	0.3650	163	0.9695	0.5542	26	0.9993	0.6804	13	**1**	**0.6936**
Adult	347	0.2290	0.1944	153	0.3804	0.2429	17	**0.7826**	**0.3281**	19	0.7276	0.3033
CC	474	0.1739	0.1693	425	0.1735	0.1692	28	**0.5586**	**0.2912**	20	0.5703	0.2798
German	361	0.2127	0.1852	210	0.2268	0.2045	28	**0.6597**	**0.3010**	25	0.5733	0.2708
LSAC	122	0.4678	0.2631	135	0.4433	0.2600	10	**0.8297**	**0.3277**	19	0.7998	0.3204
Pima	310	0.2002	0.1813	92	0.3314	0.2232	14	**0.6990**	**0.3177**	13	0.6192	0.2752

defined as usual as

$$f1(j) = 2\frac{\text{precision}(j) \cdot \text{recall}(j)}{\text{precision}(j) + \text{recall}(j)}$$

and overall $f1$ is then the average over sensitive categories, $f1 = \frac{1}{S}\sum_{j=0}^{S-1} f1(j)$. In Appendix E we show that $f1$ tends to $1/S$ as the scale grows, i.e. $\text{Perf}(\infty) = 1/S$.

For any visualization, we locate the fairness scale threshold k_{fair}, i.e. smallest k such that f1 of kNN sensitive attribute prediction is near random guessing at all $k' \geq k$. We then evaluate DR performance at fair scales: at k or all $k' \geq k$.

DR performance measure. We could evaluate DR performance at the chosen scales by any suitable measure. The aim is to evaluate how much high-dimensional data properties are preserved in DR output at the fair scales. For neighbor embedding methods it would be tempting to simply evaluate neighbor retrieval performance; however, to avoid favoring methods like NeRV designed for neighbor retrieval, we use an indirect measure: predicting high-dimensional clusters. We cluster the high-dimensional data excluding the sensitive attribute (here by k-means to 6 clusters for real data; and using the nonsensitive available clustering for synthetic data) and evaluate kNN performance of predicting the high-dimensional cluster of a point from its neighbors in DR output, by the f1 measure. We report the f1 at the fairness scale threshold k ("$f1_k$") and the average ("$f1_{\text{Avg}}$") of f1 over all scales $N > k' \geq k$, where the number of samples N yields an upper limit of kNN neighborhood size. We use an independent run of clustering on each training set and test set, to avoid overfitting parameters to a specific clustering. For more detail, Appendix H shows plots of f1 of the clusters and f1 of sensitive attribute values at all scales.

4.4. Choice of Parameters

All the methods involve neighborhood scales σ_i; we set them by an effective neighborhood size as in (Venna et al., 2010), by 20 neighbors if the minority sensitive category has over 100 points and 10 neighbors otherwise. Other hyperparameters (for ct-SNE, a neighborhood perturbation parameter; for Fair-NeRV and Fair-t-NeRV, ω and tradeoff parameters γ, β,$\tau^{\in}$,$\tau^{\notin}$) are fitted to each training set. For all methods, hyperparameters are chosen to maximize training-set performance $f1_{\text{Avg}}$ chosen over 5 rounds of uniform sampling of 3600 hyperparameter sets in the hyperparameter spaces of each method with feasible ranges for each hyperparameter (see Appendix D). Note that each method optimizes its hyperparameters over the same number of parameter combinations to keep comparison fair. The methods are then run with the chosen hyperparameters on the held-out test data and are evaluated by the $f1_k$ and $f1_{\text{Avg}}$ scores.

Our hyperparameter search takes fairness into account. Training-set performance is evaluated by $f1_{\text{Avg}}$, average f1 of cluster retrieval over the fair scales defined by fairness of sensitive attributes. Thus the search optimizes fairness and retrieval performance simultaneously. Hyperparameters yielding fairness over many scales but bad retrieval performance over them, or hyperparameters yielding fairness only over the few largest scales, will not be selected. Optimizing $f1_{\text{Avg}}$ optimizes the amount of fair scales having good retrieval: a walkthrough of this argument is in Appendix L. The $f1_{\text{Avg}}$ is based on clusters of remaining attributes; Appendix N discusses how evaluation at fair scales avoids dependency of the clustering with sensitive attributes.

For all data, the hyperparameter search suffices to pick the fairness hyperparameter β and other hyperparameters. We find optimized β values are small when sensitive attribute values have overall even proportions as in Syn (0.04) and CC data (0.05). For other data β is from 0.17 (Pima) to 0.46 (Adult). Impact of σ and λ is like in Venna et al. (2010).

4.5. Results

Quantitative measures. Table 1 reports $f1_k$ and $f1_{\text{Avg}}$ scores of different embeddings for test data sets with best

parameters obtained from training data sets. It shows how much high dimensional structure is visible at fair scales of inspection; the more fair an embedding is, and the more high-dimensional data structure is shown at fair scales, the higher the scores will be. Note that for all methods $f1_k$ will tend to report higher values than $f1_{\text{Avg}}$ as the latter is an average over multiple scales with largest scales having broad neighborhoods. Best scores on each data set are bolded.

In Table 1, for all data sets our proposed new methods Fair-NeRV and Fair-t-NeRV strongly outperform the comparison methods t-SNE and ct-SNE. Fair-NeRV/Fair-t-NeRV has smaller fairness scale threshold than tsne and ct-sne for all datasets. On the real data sets, Fair-NeRV attains the highest values of both $f1_k$ and $f1_{\text{Avg}}$, and Fair-t-NeRV has second highest values. On synthetic data the t-distributed Fair-t-NeRV has best results and Fair-NeRV is second best. Among comparison methods, ct-SNE performs reasonably especially on synthetic data, but on real data sets the score difference to Fair-NeRV and Fair-t-NeRV remains large; ct-SNE outperforms basic t-SNE on four data sets but not on the CC and LSAC data sets. Overall, Fair-NeRV and Fair-t-NeRV consistently give the best results for fair DR.

Plots. We show test data embeddings for four data sets Syn (Fig. 1) and Adult, CC, and German (Fig. 2); plots for LSAC and Pima are in Appendix G. In each figure, subplots a/a', b/b', c/c', d/d' are embeddings by t-SNE, ct-SNE, Fair-NeRV, Fair-t-NeRV, colored by sensitive attribute values/high-dimensional clusters. Fair-NeRV and Fair-t-NeRV yield fair embeddings that are informative of high-dimensional structure, outperforming t-SNE and ct-SNE.

5. Conclusions and Discussion

We introduced a fair nonlinear dimensionality reduction (DR) method Fair-NeRV and its variant Fair-t-NeRV. They optimize an information retrieval objective, avoiding errors in retrieving neighbors and retrieving a desired uninformative distribution for sensitive attributes, with a tradeoff parameter for whether neighbor preservation or fairness is more important. They find DR outputs fair to sensitive groups (e.g. groups by gender, age, religion, education) that still show high-dimensional data structure well, outperforming previous nonlinear and linear DR. The new methods can be used alongside traditional DR, to explore which variation is prominent in data and which variation can be represented without dependence on sensitive attributes.

Our cost function terms have direct information retrieval interpretations. For the neighbor retrieval term (NeRV cost function) see (Venna & Kaski, 2007); for the fairness cost the connection to information retrieval is shown in Appendix B. This is unlike e.g. ct-SNE which is only mathematically/algorithmically motivated in terms of how fairness is incorporated into its cost function. We also proposed measures $f1_k$ and $f1_{\text{Avg}}$ for fair DR quality. Our scale-based quality criterion $f1_{\text{Avg}}$ gives insight into the scales at which fairness is achieved, and allows comparison of methods.

Our method uses sensitive attributes to ensure fairness of exploration, in this sense they act as side information rather than supervision. (Supervised DR using sensitive attributes as prediction targets would do the opposite: it would arrange low-dimensional points to reveal sensitive attribute values.)

Fairness in Exploratory Data Analysis (EDA) is our primary aim. In EDA there is no target attribute to be predicted, and formulating the notion of fairness and optimizing it has been a challenge, especially for nonlinear DR where e.g. concepts of null spaces (Ravfogel et al., 2020) are not feasible. We define an objective for fair nonlinear DR based on neighbor retrieval, resulting in a task-based fairness measure suitable in EDA. Fair-NeRV visualizations complement baseline unfair visualizations in EDA, and reveal what structure of data can be visualized without revealing sensitive attributes. In experiments Fair-NeRV and Fair-t-NeRV consistently showed detailed structure of data sets; the detail they show is on an equal level as in t-SNE, thus they are usable for fair EDA. The good results of our unsupervised quality measure with clusters also suggest our results are good for EDA.

We introduced the first neighbor embedding solution designed for fair nonlinear DR. Our main comparisons were to nonlinear DR; our method yielded the best results for fair nonlinear DR in experiments. Approaches have been proposed for linear fair DR: benefit of fair nonlinear DR vs. fair linear DR is generally the same as that of nonlinear DR vs. linear DR; in Appendix O we discuss the benefits. In an experiment (Appendix K) our method outperformed a representative linear fair DR method. Thus our method outperformed both nonlinear and linear comparable DR.

We list a few future work topics. **1)** Fair-NeRV is intended for EDA, but since it yields good results for exploring data, its outputs might be usable as input features for downstream classification with respect to some prediction target. **2)** We assume a categorical sensitive attribute; extension to continuous attributes is future work. **3)** The scale-dependent sensitive attribute predictor is defined very generally but instantiated as KNN. The concept of scale/smoothing/regularization applies to other predictors too, but this deserves follow-up studies. **4)** Fair EDA case studies with domain experts.

Software. Software (Matlab & C++) and data are available at `https://github.com/wenxu-fi/Fair-NeRV`.

Acknowledgements

The work was supported by Academy of Finland decision 327352. We thank Kalervo Järvelin for insightful comments.

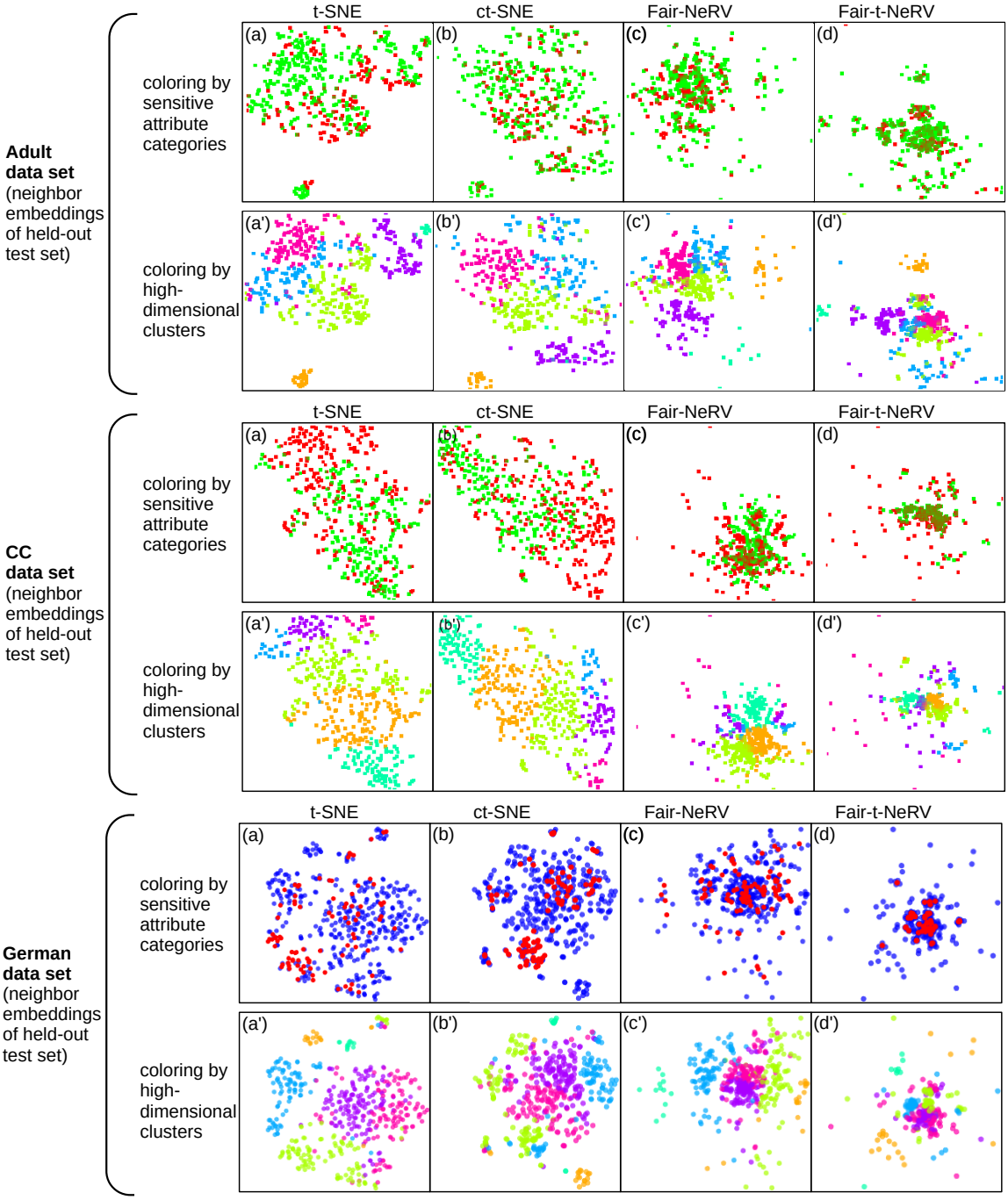

Figure 2. Test data embedings. **Top:** Adult test data. t-SNE (a) reveals sensitive information through concentrations of majority/minority values (male, green; female, red points). ct-SNE (b) mixes gender information somewhat but not fully. Fair-NeRV (c) mixes the gender information quite well and still shows nonsensitive clusters well (c'). Fair-t-NeRV (d) also mixes gender information and retains non-sensitive structure (d'). **Middle:** CC test data. t-SNE (a) reveals sensitive information by clear red and green concentrated areas at upper left and lower right. ct-SNE (b) fails similarly, green and red concentrated at top left and lower right. Fair-NeRV (c) succeeds in mixing sensitive attributes and still shows high-dimensional clusters well (c'). Fair-t-NeRV (d, d') similarly works well. **Bottom:** German test data. t-SNE and ct-SNE (a, b) reveal sensitive information in clear clusters dominated by an age group (red and blue). Fair-NeRV (c) mixes sensitive information evenly and shows high-dimensional clusters well (c'). Fair-t-NeRV (d, d') works almost as well as Fair-NeRV and still outperforms t-SNE and ct-SNE.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Assel, H. V., Espinasse, T., Chiquet, J., and Picard, F. A probabilistic graph coupling view of dimension reduction. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=tNXumks8yHv.

Colange, B., Peltonen, J., Aupetit, M., Dutykh, D., and Lespinats, S. Steering distortions to preserve classes and neighbors in supervised dimensionality reduction. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, pp. 13214–13225. Curran Associates Inc., 2020.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Hinton, G. E. and Roweis, S. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15, pp. 833–840. MIT Press, 2002. URL https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

Kamani, M. M., Haddadpour, F., Forsati, R., and Mahdavi, M. Efficient fair principal component analysis. *Mach Learn*, 111:3671–3702, 2022. doi: 10.1007/s10994-021-06100-9.

Kang, B., Garíca Garíca, D., Lijffijt, J., Santos-Rodríoguez, R., and Bie, T. D. Conditional t-SNE: more informative t-SNE embeddings. *Mach Learn*, 110:2905–2940, 2021. doi: 10.1007/s10994-020-05917-0.

Lee, J., Kim, G., Olfat, M., Hasegawa-Johnson, M., and Yoo, C. D. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7363–7371, Jun. 2022. doi: 10.1609/aaai.v36i7.20699.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018.

Olfat, M. and Aswani, A. Convex formulations for fair principal component analysis. In *AAAI'19*, volume 33, pp. 663–670, 2019.

Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://doi.org/10.1080/14786440109462720.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647.

Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. D. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Samadi, S., Tantipongpipat, U., Morgenstern, J., Singh, M., and Vempala, S. The price of fair PCA: one extra dimension. In *NIPS'18:Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10999–11010, NY,USA, 2018. Curran Associates Inc.

Tang, J., Liu, J., Zhang, M., and Mei, Q. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 287–297, International World Wide Web Conferences Steering Committee, 2016. Geneva,Switzerland. doi: 10.1145/2872427.2883041.

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

Venna, J. and Kaski, S. Nonlinear dimensionality reduction as information retrieval. In Meila, M. and Shen, X. (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 572–579, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL https://proceedings.mlr.press/v2/venna07a.html.

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

A. Gradient of the Cost Function

In these appendices we provide further details of the gradient of the cost function (Appendix A), information retrieval interpretation of the fairness cost (Appendix B), computational complexity (Appendix C), hyperparameter ranges (Appendix D), definition of fair evaluation scales (Appendix E), and drawing of the plots (Appendix F). We also show plots for data sets LSAC and Pima that were not shown in the main paper (Appendix G), plots of f1 scores over different scales (Appendix H), two comparisons to restricted special cases of our method (Appendix I and Appendix J), a comparison to the MBF-PCA method (Appendix K). We further provide discussions of the motivation of our hyperparameter search procedure (Appendix L), connection of our fairness cost to demographic parity (Appendix M), avoiding unfairness in the cluster-based performance evaluation (Appendix N), and reasons why we investigate nonlinear approaches for fair DR (Appendix O).

First, we provide the gradients for both terms of the cost function $C_{FairNeRV}$ as follows. The gradient of the fairness cost term $C_{Fairness}$ is

$$\frac{\partial C_{fairness}}{\partial y_i} = \sum_{j\neq i} \frac{\gamma}{N} \cdot \left[\frac{2}{\sigma_i^2} \left(1 - \frac{\rho_{is_j}}{r_{is_j}}\right) q_{ij} + \frac{2}{\sigma_j^2} \left(1 - \frac{\rho_{js_i}}{r_{js_i}}\right) q_{ji} \right] \cdot (y_j - y_i)$$
$$+ \sum_{j\neq i} \frac{1-\gamma}{N} \left[\frac{2}{\sigma_i^2} \left(\log \frac{r_{is_j}}{\rho_{is_j}} - D_{KL}(r_i, \rho_i) \right) q_{ij} + \frac{2}{\sigma_j^2} \left(\log \frac{r_{js_i}}{\rho_{jS_i}} - D_{KL}(r_j, \rho_j) \right) q_{ji} \right] \cdot (y_j - y_i) \ .$$

Denote $N_{ij}^{\in} = \tau^{\in}(q_{ij} - p_{ij}) + (1-\tau^{\in}) q_{ij} \log \frac{q_{ij}}{p_{ij}}$, $N_{ij}^{\notin} = \tau^{\notin}(q_{ij} - p_{ij}) + (1-\tau^{\notin}) q_{ij} \log \frac{q_{ij}}{p_{ij}}$, $G_i^{\in} = \sum_{j\in S_i^{\in}} N_{ij}^{\in}$ and $G_i^{\notin} = \sum_{j\in S_i^{\notin}} N_{ij}^{\notin}$. The gradient of the neighbor retrieval cost term C_{CNeRV} is

$$\frac{\partial C_{CNeRV}}{\partial y_i} = \sum_{j\neq i} \left(\frac{2}{\sigma_i^2} G_i^{\in} q_{ij} + \frac{2}{\sigma_j^2} G_j^{\in} q_{ji} \right) \cdot (y_i - y_j) - \sum_{j\in S_i^{\in}} \left(\frac{2}{\sigma_i^2} N_{ij}^{\in} + \frac{2}{\sigma_j^2} N_{ji}^{\in} \right) \cdot (y_i - y_j)$$
$$+ \sum_{j\neq i} \left(\frac{2}{\sigma_i^2} G_i^{\notin} q_{ij} + \frac{2}{\sigma_j^2} G_j^{\notin} q_{ji} \right) \cdot (y_i - y_j) - \sum_{j\in S_i^{\notin}} \left(\frac{2}{\sigma_i^2} N_{ij}^{\notin} + \frac{2}{\sigma_j^2} N_{ji}^{\notin} \right) \cdot (y_i - y_j) \ .$$

B. Information Retrieval Interpretation

We show the fairness cost has an information retrieval interpretation.

Theorem 3.1. (restated from the main paper). The cost $C_{Fairness}$ in Eq. (4) is an average information retrieval cost of misses and false neighbors. In detail, under simplifying assumptions,

$$C_{Fairness} \approx \frac{const.}{N} \sum_i \big(\gamma(1 - recall(i)) + (1-\gamma)(1 - precision(i)) \big)$$

where $recall$ and $precision$ are the typical information retrieval measures, in the task of retrieving the desired distribution of sensitive attribute values for each point from the low-dimensional display. More precisely, recall is the proportion of true positives out of all positives, and precision is the proportion of true positives out of all retrieved items, where the sets of true positives, all positives, and retrieved items will be defined in the proof below.

Proof. The fairness cost $C_{Fairness}$ is a weighted average of several Kullback-Leibler divergences of two kinds, $D_{KL}(\rho_i, r_i)$ and $D_{KL}(r_i, \rho_i)$. We first show each such individual divergence has an information retrieval interpretation in terms of information retrieval costs of misses and false neighbors, by a proof analogous to one used for neighbor retrieval costs in (Venna et al., 2010). We then discuss the interpretation of the full fairness cost.

Consider the distribution of sensitive attributes around a data point i in the low-dimensional space. Suppose the desired distribution of sensitive attributes ρ_i has uniform high probabilities for a subset of the sensitive attribute values (denote the subset as A_i and the number of sensitive attribute categories in it as $L_i = |A_i|$) and very low uniform probabilities for all other $S - L_i$ classes, where S is the total number of possible sensitive attribute values. The probability of a sensitive attribute category can then be written as $\rho_i(s) = (1-\kappa)/L_i$ for categories $s \in A_i$ and $\rho_i(s) = \kappa/(S - L_i)$ for categories

$s \notin A_i$, where κ is a very small positive number. Similarly, suppose that the low-dimensional distribution r_i has uniform high probabilities for a subset of categories (denote the subset as B_i and the number of sensitive attribute categories in it as $M_i = |B_i|$) and very low probabilities for all other $S - M_i$ categories. The probability of a sensitive attribute category can then be written as $r_i(s) = (1-\kappa)/M_i$ for categories $s \in B_i$ and $r_i(s) = \kappa/(S - M_i)$ for categories $s \notin B_i$. Note that the sets A_i and B_i may differ from each other including their sizes, and each point i has its own sets A_i and B_i.

Then the divergence $D_{KL}(\rho_i, r_i)$ becomes

$$
\begin{aligned}
D_{KL}(\rho_i, r_i) &= \sum_{s=1}^{S} \rho_i(s) \log \frac{\rho_i(s)}{r_i(s)} \\
&= \sum_{s \in A_i \cap B_i} \rho_i(s) \log \frac{\rho_i(s)}{r_i(s)} + \sum_{s \in A_i \setminus B_i} \rho_i(s) \log \frac{\rho_i(s)}{r_i(s)} + \sum_{s \in B_i \setminus A_i} \rho_i(s) \log \frac{\rho_i(s)}{r_i(s)} + \sum_{s \notin B_i \cup A_i} \rho_i(s) \log \frac{\rho_i(s)}{r_i(s)} \\
&= \sum_{s \in A_i \cap B_i} \frac{1-\kappa}{L_i} \log \frac{(1-\kappa)/L_i}{(1-\kappa)/M_i} + \sum_{s \in A_i \setminus B_i} \frac{1-\kappa}{L_i} \log \frac{(1-\kappa)/L_i}{\kappa/(S-M_i)} \\
&\quad + \sum_{s \in B_i \setminus A_i} \frac{\kappa}{S-L_i} \log \frac{\kappa/(S-L_i)}{(1-\kappa)/M_i} + \sum_{s \notin B_i \cup A_i} \frac{\kappa}{S-L_i} \log \frac{\kappa/(S-L_i)}{\kappa/(S-M_i)} \\
&= |A_i \cap B_i| \frac{1-\kappa}{L_i} \log \frac{M_i}{L_i} + |A_i \setminus B_i| \frac{1-\kappa}{L_i} \left(\log \frac{1-\kappa}{\kappa} + \log \frac{S-M_i}{L_i} \right) \\
&\quad + |B_i \setminus A_i| \frac{\kappa}{S-L_i} \left(\log \frac{\kappa}{1-\kappa} + \log \frac{M_i}{S-L_i} \right) + (S - |A_i \cup B_i|) \frac{\kappa}{S-L_i} \log \frac{S-M_i}{S-L_i} \qquad (6)
\end{aligned}
$$

and because the terms $\log[(1-\kappa)/\kappa]$ dominate other terms when κ is small, the above further simplifies to

$$
D_{KL}(\rho_i, r_i) \;\approx\; \left(|A_i \setminus B_i| \frac{1-\kappa}{L_i} \;-\; |B_i \setminus A_i| \frac{\kappa}{S-L_i} \right) \log \frac{1-\kappa}{\kappa} \;\approx\; \frac{|A_i \setminus B_i|}{L_i} (1-\kappa) \log \frac{1-\kappa}{\kappa} \;=\; \frac{N_{MISS,i}}{L_i} C \qquad (7)
$$

where $N_{MISS,i} = |A_i \setminus B_i|$ is the number of missed sensitive attribute values that had high probabilities in the desired distribution but low probabilities in the actual low-dimensional distribution around point i, and C is a constant depending only on κ. Further, since recall of a query is the rate of true positives $N_{TP,i} = |A_i \cap B_i|$ out of all positives $N_{POS,i} = |A_i| = L_i$, we have $recall(i) = N_{TP,i}/N_{POS,i} = |A_i \cap B_i|/L_i = (|A_i| - |A_i \setminus B_i|)/L_i = 1 - N_{MISS,i}/L_i$, and therefore $N_{MISS,i}/L_i = 1 - recall(i)$. Thus the dominating term of the divergence is directly proportional to $1 - recall(i)$. Minimizing the divergence thus maximizes recall of retrieving categories of the desired sensitive attribute distribution. The exact relationship holds under the assumptions at the start of the proof, more generally with free-form distributions ρ_i and r_i minimizing the divergence can be seen as maximizing a generalization of recall.

Similarly, for the divergence $D_{KL}(r_i, \rho_i)$ we have

$$
\begin{aligned}
D_{KL}(r_i, \rho_i) &= \sum_{s=1}^{S} r_i(s) \log \frac{r_i(s)}{\rho_i(s)} \\
&= \sum_{s \in B_i \cap A_i} \frac{1-\kappa}{M_i} \log \frac{(1-\kappa)/M_i}{(1-\kappa)/L_i} + \sum_{s \in B_i \setminus A_i} \frac{1-\kappa}{M_i} \log \frac{(1-\kappa)/M_i}{\kappa/(S-L_i)} \\
&\quad + \sum_{s \in A_i \setminus B_i} \frac{\kappa}{S-M_i} \log \frac{\kappa/(S-M_i)}{(1-\kappa)/L_i} + \sum_{s \notin B_i \cup A_i} \frac{\kappa}{S-M_i} \log \frac{\kappa/(S-M_i)}{\kappa/(S-L_i)} \\
&= |B_i \cap A_i| \frac{1-\kappa}{M_i} \log \frac{L_i}{M_i} + |B_i \setminus A_i| \frac{1-\kappa}{M_i} \left(\log \frac{1-\kappa}{\kappa} + \log \frac{S-L_i}{M_i} \right) \\
&\quad + |A_i \setminus B_i| \frac{\kappa}{S-M_i} \left(\log \frac{\kappa}{1-\kappa} + \log \frac{L_i}{S-M_i} \right) + (S - |B_i \cup A_i|) \frac{\kappa}{S-M_i} \log \frac{S-L_i}{S-M_i} \qquad (8)
\end{aligned}
$$

and because the terms $\log[(1-\kappa)/\kappa]$ again dominate this simplifies to

$$D_{KL}(r_i,\rho_i) \;\approx\; \left(|B_i\backslash A_i|\frac{1-\kappa}{M_i} \;-\; |A_i\backslash B_i|\frac{\kappa}{S-M_i}\right)\log\frac{1-\kappa}{\kappa} \;\approx\; \frac{|B_i\backslash A_i|}{M_i}(1\;-\;\kappa)\log\frac{1-\kappa}{\kappa} \;=\; \frac{N_{FP,i}}{M_i}C \quad (9)$$

where $N_{FP,i} = |B_i\backslash A_i|$ is the number of false positives, sensitive attribute values that had low probabilities in the desired distribution but high probabilities in the actual low-dimensional distribution around point i, and the constant C again only depends on κ. Since precision of a query is the rate of true positives $N_{TP,i}$ out of all retrieved items, i.e. sensitive values with high low-dimensional probability, denoted $N_{RETR,i} = |B_i| = M_i$, we have $precision(i) = N_{TP,i}/N_{RETR,i} = |A_i \cap B_i|/M_i = (|B_i| - |B_i\backslash A_i|)/M_i = 1 - N_{FP,i}/M_i$, and therefore $N_{FP,i}/M_i = 1 - precision(i)$. The dominating term of the divergence is thus proportional to $1 - precision(i)$ and minimizing the divergence maximizes precision of retrieving categories of the desired sensitive attribute distribution. Again, the exact relationship holds under the assumptions at the start of the proof, and more generally with free-form distributions ρ_i and r_i minimizing the divergence can be seen as maximizing a generalization of precision.

Given the information retrieval formulation of each divergence, the total fairness cost can then be written as

$$C_{Fairness} = \frac{1}{N}\sum_i \Big(\gamma D_{KL}(\rho_i,r_i) + (1-\gamma)D_{KL}(r_i,\rho_i)\Big)$$
$$\approx \frac{C}{N}\sum_i \Big(\gamma(1-recall(i)) + (1-\gamma)(1-precision(i))\Big), \quad (10)$$

which is (up to the constant factor C) the averaged cost of information retrieval errors (in terms of precision and recall) when retrieving the categories of the desired sensitive category distributions for each of the N data points, and γ controls the balance between precision and recall. This completes the proof. □

C. Computational Complexity

It is easy to see that computation of the cost function and gradient of Fair-NeRV and Fair-t-NeRV has complexity $\mathcal{O}(n^2)$ with respect to the number of data points n, since computation can be organized to compute terms such as divergences in a first step which are then used for final computation of the cost and gradients. Therefore the complexity of the new methods is the same as that of other neighbor embedding methods such as t-SNE, ct-SNE, NeRV, and ClassNeRV. Also, like several other neighbor embedding methods, our method can be optimized with a Barnes-Hut based approximation for time complexity $\mathcal{O}(n \log n)$. FIt-SNE style reformulation is future work. In this work we focused on the idea rather than the efficiency.

An individual run of Fair-NeRV or Fair-t-NeRV on one of the experiment data sets is very fast, taking less than a minute. For the experiments with hyperparameter search, for all methods we ran different hyperparameter combinations in parallel on a computing cluster, before selecting the final hyperparameter combinations.

D. Hyperparameter Ranges

As mentioned in Section 4.4 of the main paper, hyperparameters of methods are chosen to maximize training-set performance $f1_{\text{Avg}}$ chosen over 5 rounds of uniform sampling of 3600 hyperparameter sets in the hyperparameter spaces of each method with feasible ranges. In our experiments, we set the ranges for Fair-NeRV and Fair-t-NeRV parameters to be $\tau^{\in} \in [0,1]$, $\tau^{\notin} \in (\tau^{\in},1]$, $\beta \in [0,1]$, $\gamma \in [0,1]$, and $\omega \in [0.5, 0.99]$. The ct-SNE method has a hyperparamer β' whose range is $[10^{-7},1]$. Out of the 5 rounds of uniform sampling of ranges for Fair-NeRV/tFair-NeRV (log-uniform sampling of the range for the ct-SNE hyperparameter), the hyperparameter combination having the highest $f1_{\text{Avg}}$ is chosen and is then used for evaluation on the test set.

E. Fair Scale Definition

As mentioned in the main paper, an evaluation scale is considered fair if the f1 score of sensitive attribute classification, here denoted $f1^{(\text{Sensitive})}$, is close to that of random guessing. We now derive the random guessing performance.

Theorem E1. With the soft precision, recall, and f1 definitions in the main paper, performance of random guessing according to overall category proportions equals $1/S$ where S is the number of sensitive categories. Moreover, f1 of soft kNN classification of sensitive attribute values tends to f1 of such random guessing as scale grows.

Proof. It is easy to show that f1 of soft kNN classification tends to random guessing as scale grows: as the scale k in kNN grows towards using all neighbors, neighbor proportion $\epsilon_{i,j}$ of any sensitive attribute category j around each point i tends towards the overall proportion of the category in the data set, denoted $u(j)$. Then it is easy to show that soft precision, soft recall, and the corresponding f1 measure all tend to $u(j)$. Denote the set of points having sensitive value j by Z_j.

- For precision, we have $precision(j) = (\sum_{i\in Z_j} \epsilon_{i,j})/(\sum_{i'} \epsilon_{i',j}) \to (\sum_{i\in Z_j} u(j))/(\sum_{i'} u(j)) = u(j)$.
- For recall, we similarly have $recall(j) = (\sum_{i\in Z_j} \epsilon_{i,j})/|Z_j| \to (\sum_{i\in Z_j} u(j))/|Z_j| = u(j)$.
- For the f1 score we have $f1^{(\text{Sensitive})}(j) = 2\frac{precision(j)\cdot recall(j)}{precision(j)+recall(j)} \to 2\frac{u(j)\cdot u(j)}{u(j)+u(j)} = u(j)$.

Therefore the average value of one of these measures over all S sensitive categories simply tends to the reciprocal of the number of sensitive categories: for the f1 score we have $f1^{(\text{Sensitive})} = \frac{1}{S}\sum_{j=1}^{S} f1^{(\text{Sensitive})}(j) \to \frac{1}{S}\sum_{j=1}^{S} u(j) = \frac{1}{S}$, and precision and recall yield the same result. This concludes the proof. □

We thus call a scale k fair if the f1 at that scale, denoted $f1_k^{(\text{Sensitive})}$, is close to random guessing so that $|f1_k^{(\text{Sensitive})} - \frac{1}{S}| \leq \epsilon$, where ϵ is a very small allowed difference. In experiments we set the allowed difference to be a small fraction, here 1%, of the range between random guessing and perfect f1 score: $\epsilon = 0.01 \cdot (1 - \frac{1}{S})$.

F. Plot Drawing Details

Since we deal with sensitive attributes having minority categories, in visualizations it is important not to obscure sensitive information (and hinder visual assessment of plot quality) with overplotting issues. Therefore in experiments, for all methods the images colored by sensitive attribute categories are drawn to best highlight the sensitive distribution as follows.

If a data set has a small sensitive category (with less than 100 points, i.e., about 20% of all points in a plot of 500 points), we plot the categories in order of size largest first to ensure minority categories are not overplotted: this is done for the German, LSAC and Pima data sets. Otherwise, i.e., when the sensitive categories have a fairly even overall distribution, we use transparency-based color mixing for the sensitive category colored plots, to ensure all sensitive categories are shown where they occur: this is done for the Syn, Adult, and CC data sets.

G. Additional Visualizations

We show visualizations of the Law School (LSAC; Figure 3) and Pima (Figure 4) data sets here, in a similar format to Figure 2 of the main paper. As discussed in the main paper, the new Fair-NeRV and Fair-t-NeRV methods clearly outperform the comparison methods, yielding visualizations that are fair with respect to sensitive attributes but are still able to be informative about high-dimensional data structure, as seen here from high-dimensional clusters.

H. Plots of f1 Score

Figures 5 and 6 show plots of f1 score of sensitive attribute values and f1 of high-dimensional clusters over different neighborhood scales k (horizontal axis in the figures), for two of the data sets. In these figures, the desired result for fair visualization is that 1) the f1 for sensitive attribute values would remain low over as many scales as possible, so that the low-dimensional location of a point cannot be used to accurately retrieve its sensitive attribute value, and 2) the f1 for high-dimensional clusters would remain high over as many scales as possible, especially those scales where the sensitive-attribute f1 is low, so that the low-dimensional visualization remains informative about the high-dimensional organization of data.

The figures show that Fair-NeRV/ Fair-t-NeRV achieves clearly smaller f1 scores for sensitive attribute values over the neighborhood scales than other methods. Therefore, Fair-NeRV/Fair-t-NeRV also achieves a smaller fairness scale threshold than other methods, i.e., it reaches sooner a scale after which the f1 of sensitive attribute values remains low. Also, at the

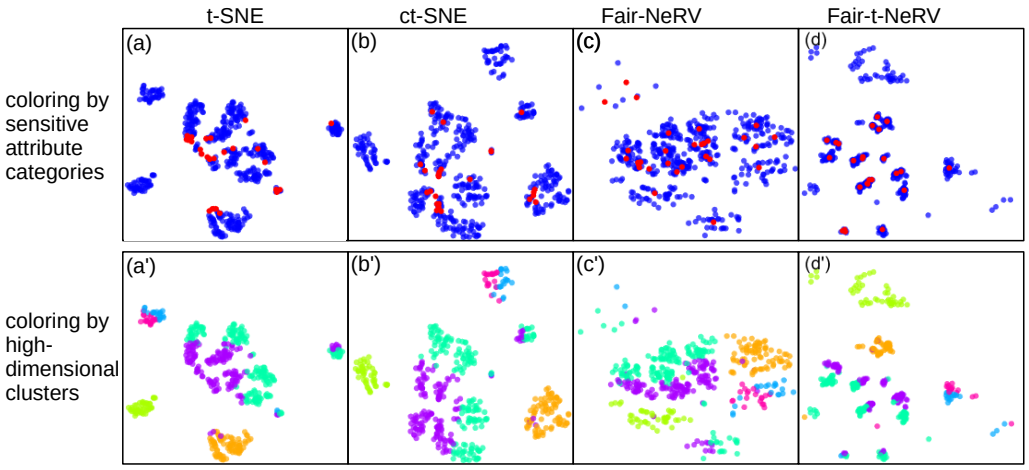

Figure 3. Dimensionality reduction results for the LSAC data set. Embeddings by t-SNE (a, a') and ct-SNE (b, b') clearly reveal sensitive information through concentrations of the minority and majority sensitive categories (applicant's race, red dots: non-white, blue dots: white). In contrast, the Fair-NeRV output (c, c') has an even distribution of the sensitive attribute values throughout the embedding, while still showing high dimensional data organization well, as seen from the coloring by high-dimensional clusters. The Fair-t-NeRV version (d, d') similarly works well.

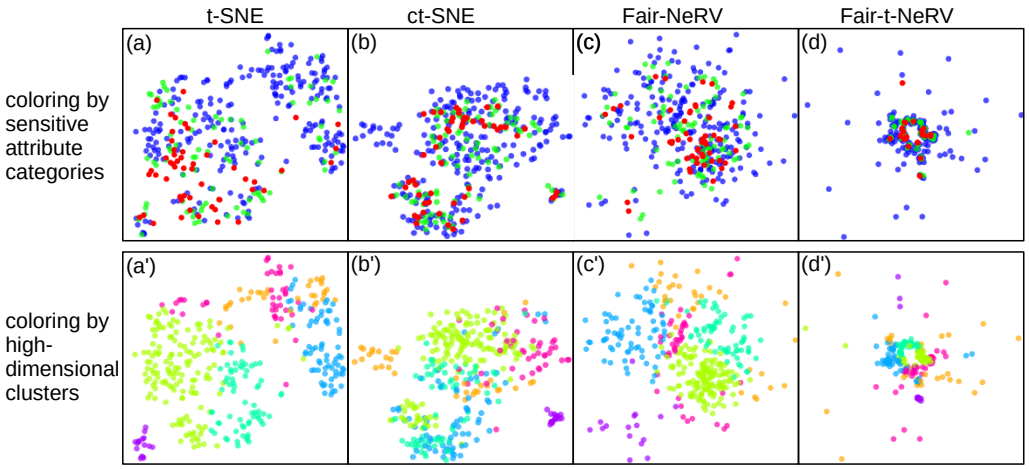

Figure 4. Dimensionality reduction results for the Pima data set. Outputs of t-SNE (a, a') and ct-SNE (b, b') show clusters having strong concentrations of particular sensitive groups (BMI categories, normal weight: red; overweight: green; obese: blue), thus the low-dimensional organization reveals sensitive attributes. Our method Fair-NeRV (c, c') mixes sensitive groups better than t-SNE and ct-SNE while still showing high dimensional data organization well, as seen from the coloring by high-dimensional clusters. The Fair-t-NeRV version (d, d') also works well.

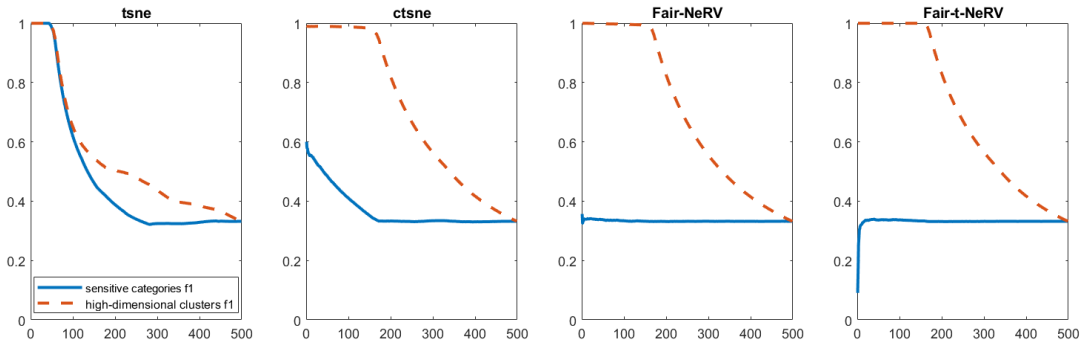

Figure 5. Graph of f1 of sensitive attribute values (sensitive categories) and high-dimensional non-sensitive clusters for syn data.

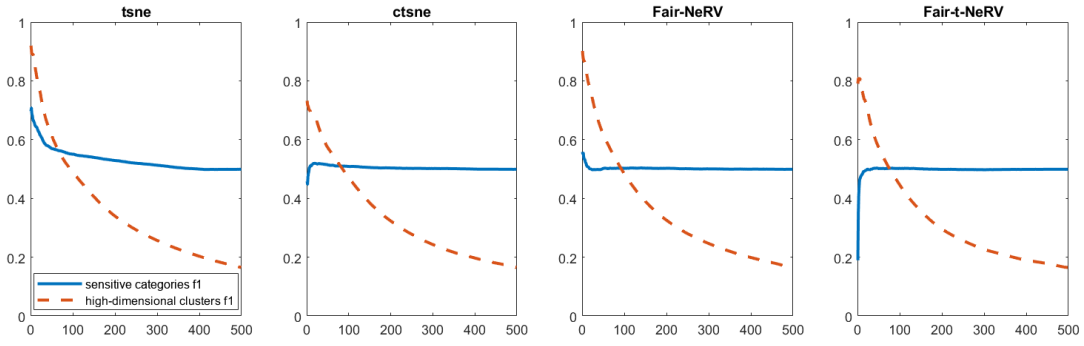

Figure 6. Graph of f1 of sensitive attribute values (sensitive categories) and high dimensional clusters for adult data

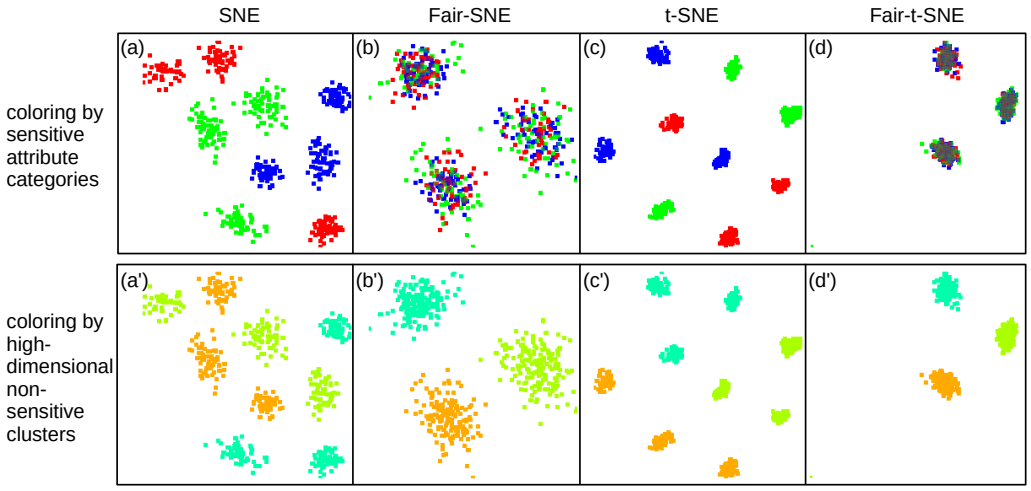

Figure 7. Dimensionality reduction results for SyntheticData. Outputs of SNE (a, a'), Fair-SNE (b, b'), t-SNE (c,c'), Fair-t-SNE (d,d'). Fair-SNE/Fair-t-SNE mixes sensitive groups quite well and remain high dimensional clusters.

fairness scale threshold, our method will have higher f1 of high-dimensional clusters. The graphs for all other real data sets show similar behavior as the graphs of syn and adult data shown here.

I. Special Case

In general our method optimizes its internal parameters including the tradeoff parameters $\tau^{\in}$, $\tau^{\notin}$. However, it is possible to constrain $\tau^{\in} = \tau^{\notin} = 1$ in Fair-NerV and Fair-t-NeRV, which yields fair versions of SNE and t-SNE as special cases. We will denote these special cases here as Fair-SNE and Fair-t-SNE. (Note that in the t-distributed version we use conditional neighbor distributions instead of joint distributions of points and their neighbors, but the behavior will be similar to the joint case, i.e. the t-distributed version avoids a "crowding problem" as low-dimensional t-distributed neighborhoods reach far-off neighbors.)

As mentioned in the main paper, the special cases have similar performance as the corresponding full Fair-NeRV and Fair-t-NeRV. As an example, in Figure 7 we show plots of SNE, Fair-SNE, t-SNE, and Fair-t-SNE embeddings for the synthetic data.

In Figure 7 the special case results show that good results combining neighbor retrieval performance with fairness can be achieved without detailed consideration of the ClassNeRV tradeoff parameters, instead focusing only on the balance of the neighbor retrieval and the fairness cost. Therefore, the good performance of the Fair-SNE and Fair-t-SNE special cases shows that the fairness cost is the essential new part of our proposed cost function $C_{FairNeRV}$. This is as expected, as the ClassNeRV tradeoff parameters were only incorporated to further help a good balance, they do not otherwise play an essential role by themselves.

J. Comparison with ClassNeRV

In the previous appendix we examined performance of special cases that fixed ClassNeRV tradeoff parameters and thus focused on impact of the fairness cost. To further show how the fairness cost plays the essential role, we have made a full ablation study investigating the performance of the "unfair" case of using the ClassNeRV/t-ClassNeRV parts of the cost function alone. That is, we now consider the special case where the fairness cost is left out by restricting $\beta = 1$ in Fair-NeRV/Fair-t-NeRV. Note that in this special case, although the fairness cost is left out, the ClassNeRV/t-ClassNeRV cost does make use of the sensitive attribute labels to define what are within-class and between-class errors, and the tradeoff hyperparameters are optimized just like all hyperparameters in ClassNeRV/t-ClassNeRV. The results are in Table 2.

Comparing Table 1 and Table 2, the results of the full method Fair-NeRV/Fair-t-NeRV in Table 1 are clearly better than the results of the ClassNeRV/t-ClassNeRV special case in Table 2. The ClassNeRV/t-ClassNeRV special case performs roughly

Table 2. Dimensionality reduction performance on test data sets for ClassNeRV)

Data	ClassNeRV			t-ClassNeRV		
	k	$f1_k$	$f1_{\text{Avg}}$	k	$f1_k$	$f1_{\text{Avg}}$
Syn	374	0.4427	0.3822	333	0.5057	0.4063
Adult	241	0.2739	0.2109	174	0.2986	0.2157
CC	470	0.1755	0.1701	435	0.1819	0.1731
German	114	0.4148	0.2552	97	0.3495	0.2171
LSAC	139	0.4135	0.2538	55	0.5688	0.2736
Pima	189	0.2793	0.2085	160	0.3032	0.2124

similarly to t-SNE and ct-SNE, i.e., its results are unfair and it is clearly worse than Fair-NeRV/Fair-t-NeRV. Therefore, this ablation study provided further evidence of the importance of the fairness cost function in addition to the results of the previous appendix, showing that the ClassNeRV part is not enough for good fair performance by itself.

K. Comparison to MBF-PCA

In this appendix we provide a comparison of our method to the linear fair dimensionality reduction method MBF-PCA (Lee et al., 2022). We first provide a detailed discussion of theoretical differences between our method and MBF-PCA. We next provide experimental results for MBF-PCA, showing our method outperforms it.

Discussion of the MBF-PCA method. The method MBF-PCA used an MMD of dimensionality -reduced conditional distributions of different protected categories which is stated to generalize to demographic parity, equal opportunity, and equalized odds. The MMD definition of disrepancy, as defined in eq. (5) of the paper (Lee et al., 2022), is defined through sums and differences of direct kernel values within one distribution (that of the X_i) and another (that of the Y_j) and across them. Those kernel values are very much reliant on the scale parameter of the kernel since the terms have no normalization.

In particular, when the scale of a RBF kernel (σ) decreases towards zero, only diagonal kernel values remain nonzero, and MMD becomes close to $\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}$. And when the RBF kernel scale grows towards infinity, the MMD decreases towards zero. Therefore, the MMD approach centrally depends on the kernel scale, and can easily be biased unless it is carefully controlled. The kernel scale there is set for the needs of PCA: "For the choice of bandwidth , the median of the set of pairwise distances of the samples after vanilla PCA is considered". This choice is only reasonable since their method is a linear projection; in a nonlinear method there is no reason why the PCA-based scale would be best. Moreover, their PCA-based choice may yield a quite large (overly-smoothed) scale since the median is over all pairwise distances. As we have mentioned, overly large scales will seem fair for any embedding, thus they alone will not be good choices for a scale-dependent evaluation of embedding quality.

In contrast to MBF-PCA, in our approach all the kernel-based quantities (neighbor probabilities, sensitive attribute category probabilities) are properly normalized and do not suffer from the above problem of scale. Our neighbor retrieval probabilities and sensitive attribute category probabilities do not reduce to constants even at the smallest scales.

Experimental comparison with the MBF-PCA method. We ran MBF-PCA for all data sets that have two sensitive attribute values: Adult, German, LSAC, and CC. This is because the implementation of MBF-PCA provided by its authors is restricted to the binary case of two sensitive attribute values. We used the Demographic Parity version of MBF-PCA which is applicable to this exploratory data analysis scenario (having input features and a sensitive attribute, but no target class). We ran MBF-PCA with the authors' recommended settings from the paper of Lee et al. (2022). However, to make sure this did not yield a disadvantage for the method, we also performed a full hyperparameter search on the Adult data set, as we had done for the nonlinear methods. That result is denoted as "Adult (FHS)" below. The MBF-PCA results are clearly worse than those of Fair-NeRV and Fair-t-NeRV on all of the data sets. The full hyperparameter search on the Adult data set improved MBF-PCA results only a little, and the advantage of Fair-NeRV and Fair-t-NeRV remains clear. Thus, the results demonstrate that our nonlinear approach offers a strong advantage for fair dimensionality reduction in exploratory

Table 3. Dimensionality reduction performance on test data sets for MBF-PCA)

Data	MBF-PCA		
	k	$f1_k$	$f1_{\text{Avg}}$
Adult	128	0.2935	0.2201
Adult(FHS)	133	0.3313	0.2386
CC	140	0.2433	0.2022
German	112	0.3220	0.2342
LSAC	117	0.3103	0.2215

data analysis settings.

We have also made a tentative brief comparison to the method of Ravfogel et al. (2020), where the first result indicated our method also seemed to outperform that method. However, a more thourough comparison is left for future work.

L. Hyperparameter Search Motivation

In this appendix we provide a more detailed walkthrough of the argument in Section 4.4 of the main paper, in the paragraph starting "Our hyperparameter search takes fairness into account", showing that the hyperparameter selection criterion $f1_{\text{Avg}}$ takes into consideration both retrieval performance and fairness.

As discussed in Section 4.3, as the inspection scale grows prediction performance of predicting sensitive attributes will fall towards that of random guessing and visualizations will look "fair" at large enough scales. Therefore, fairness is easy to achieve when the inspection scale grows very large. Thus, most embeddings have fair scales at very large values of k_{fair}. However, retrieval performance at such large overly-smoothed scales is usually quite poor, because the retrieval returns too many points to get only the correct ones (here, points from same high-dimensional cluster as the center point).

If only the large scales are fair, their poor retrieval will dominate $f1_{\text{Avg}}$: the $f1_{\text{Avg}}$ will be an average of several poor retrieval performances. To get higher $f1_{\text{Avg}}$, one must get better retrieval performance values included into the average. Better values can, for real data, practically be achieved only at smaller scales (lower values of k_{fair}). Therefore, to get higher $f1_{\text{Avg}}$, one must be able to make smaller scales fair: only at the smaller scales there is a chance to get good retrieval. (Note that retrieval is not automatically good at all small scales: the embedding must be well arranged at those scales to yield good retrieval.)

As the number of small fair scales with good retrieval increases, the $f1_{\text{Avg}}$ will increase since the average is no longer dominated by the poor retrieval at overly-large scales. Therefore, optimizing $f1_{\text{Avg}}$ optimizes the amount of fair scales having good retrieval.

M. Connection to Demographic Parity

In this appendix we detail the connection of our fairness measure to demographic parity, which was mentioned in Section 3.1 of the main paper.

Our fairness cost measures whether the distribution of sensitive attributes in each low-dimensional neighborhood is close to a desired distribution. It measures for each local neighborhood whether the opportunity to be a neighbor is equal to (or close to it) for all sensitive values. In a situation where demographic parity is achieved, each individual candidate neighbor should have, a priori, equal opportunity to be a neighbor regardless of its sensitive attribute value. Therefore, for each sensitive attribute value, the total proportion of neighbors from it (total neighbor probability over the candidates) should correspond to the overall proportion of that sensitive attribute value in the data. When our desired distributions correspond to the overall proportions, our fairness cost measures whether the condition in the above bullet point holds. For each neighborhood the fairness cost measures whether the neighbor proportions correspond to the overall sensitive attribute proportions. Therefore, our fairness cost measures demographic parity. Note that if the parity (independence of the neighbor opportunity among all sensitive attribute values) is achieved in all local neighborhoods, then this can also be seen as independence between the

embedding coordinates of a point and the sensitive attribute value of its neighbors.

N. Cluster-based Performance Evaluation

As described in the main paper, the criterion $f1_{\text{Avg}}$ is computed based on a clustering of remaining attributes. In this appendix we discuss how the cluster-based criterion aims to avoid unfairness with respect to sensitive attributes.

Removing sensitive attributes from input does not suffice to ensure fairness. Removing a sensitive attribute from the input feature set is a useful first attempt towards fairness but it is not enough alone to make a DR method fair: remaining attributes can have statistical dependencies with the sensitive attribute, thus DR based on the remaining attributes could reveal information of the sensitive attributes.

Our performance measure specifically checks how much information of other attributes (other than the sensitive ones) the visualization still reveals, in terms of cluster retrieval performance, at fair inspection scales. For this purpose, the high-dimensional clusters we inspect must of course not be constructed using the sensitive attribute.

Cluster-based evaluation is appropriate when done at fair scales. If our cluster-based performance criterion is based on the remaining attributes, could dependencies between those remaining attributes and the sensitive attributes still cause bias in the cluster-based criterion? It is possible that the remaining attributes contain some dependency with the sensitive attributes, but this is not a problem for the evaluation measure: the inspection is done at fair scales, whose fairness is directly ensured with respect to the sensitive attributes.

In detail, if the clustering had dependencies with the sensitive attributes, this would mean retrieval of clusters based on the neighbors of data points could partially retrieve also information of sensitive attributes. However, if that was the case at some evaluation scale, then also the retrieval of the sensitive attributes based on the same neighbors would work better than at random. Our evaluation is designed to choose scales at which such retrieval of sensitive attributes does not succeed.

Because the evaluation is done over fair scales, the information visible from the clustering at those scales will only reveal remaining fair information of the data structure.

O. Reasons to Investigate Nonlinear Approaches for Fair DR

In general, the advantage of fair nonlinear DR over fair linear DR is the same as the advantage of nonlinear DR over linear DR. Therefore in this appendix we provide a general discussion of advantages of nonlinear DR over linear DR.

High-dimensional data may occupy a complicated topological shape in the original data space, so that a linear projection is unable to show its essential structure. Nonlinearity provides additional flexibility that allows to preserve the data structure much better. A linear projection is of course a special case of a general nonlinear mapping: if, for some data set, a mapping that corresponds to a linear or almost linear projection suffices, then the nonlinear optimization should find that mapping as an optimum of its cost function. In fact, sometimes the nonlinear methods are initialized by a PCA mapping (this is a possible initialization for all of the neighbor embedding methods compared in our paper), but the methods then continue optimization from there, so that the end result is better than a PCA result.

The benefit of nonlinearity is most crucial when the output dimensionality is small (e.g. 2- or 3-dimensional output needed for visualization), and the original dimensionality is large, therefore the essential structure of data must be preserved in a much smaller dimensionality. The classical simple examples of this advantage include manifolds like S-curves and swiss rolls, that cannot be"unfolded" by a linear projection, whereas nonlinear methods can unfold them. Venna et al. (2010) shows another simple example: reducing points on the surface of a sphere (e.g. cities on the surface of the world) onto a 2D output. Nonlinear mappings allow, for example, to create an "orange-peel world map" which is not possible by linear methods. The paper also shows the advantage nonlinear methods over PCA for multiple real high-dimensional data sets (see Figs. 4-5 in that paper for quantitative studies).

More generally, the above discussed flexibility is why nonlinear methods including neighbor embedding have become the most popular approach for data visualization in the machine learning community; in particular the t-SNE method has become a de facto standard in countless papers. This motivates the need for nonlinear solutions in fair DR as well.