Cluster Exploration Using Informative Manifold Projections

Stavros Gerolymatos[a,*]**, Xenophon Evangelopoulos**[b,c]**, Vladimir V. Gusev**[a] **and John Y. Goulermas**[a]

[a]Department of Computer Science, University of Liverpool, Liverpool, UK
[b]Department of Chemistry, University of Liverpool, Liverpool, UK
[c]Leverhulme Research Centre for Functional Materials Design, University of Liverpool, Liverpool, UK
{s.gerolymatos, evangx, vladimir.gusev, j.y.goulermas}@liverpool.ac.uk

Abstract. Dimensionality reduction (DR) is one of the key tools for the visual exploration of high-dimensional data and uncovering its cluster structure in two- or three-dimensional spaces. The vast majority of DR methods in the literature do not take into account any prior knowledge a practitioner may have regarding the dataset under consideration. We propose a novel method to generate informative embeddings which not only factor out the structure associated with different kinds of prior knowledge but also aim to reveal any remaining underlying structure. To achieve this, we employ a linear combination of two objectives: firstly, contrastive PCA that discounts the structure associated with the prior information, and secondly, kurtosis projection pursuit which ensures meaningful data separation in the obtained embeddings. We formulate this task as a manifold optimization problem and validate it empirically across a variety of datasets considering three distinct types of prior knowledge. Lastly, we provide an automated framework to perform iterative visual exploration of high-dimensional data.

1 Introduction

Data exploration focuses on identifying informative patterns to discover new insight and knowledge about a collection of data. The often high-dimensional nature of such data renders the visual exploration process intractable for the human eye, and therefore specialized data manipulation of the original samples is essential in practice. Dimensionality reduction methods have been at the forefront of this challenge [6] aiming to recover lower-dimensional embeddings of the original data that facilitate the identification of underlying data cohorts and help understand better the problem at hand.

One of the most well known dimensionality reduction approaches perhaps is principal component analysis (PCA) [15], an efficient linear method aiming to maximizing the variance along the projection vectors, which in practice appears insufficient for meaningful separation of cohorts. A variety of non-linear methods have also been proposed that conversely focus on locally preserving the structure of the data such as Isomap [37], LLE [31], t-SNE [40], UMAP [23], TriMap [3] and LargeVis [36], etc. Projection pursuit (PP) [12] defines a family of dimensionality reduction methods that can enable various embedding effects depending on a suitably selected criterion. The kurtosis index [9] is one specific PP example that specializes in identifying "interesting" projections. Its minimization partic-

ularly penalizes the normality of the data distribution, promoting thus more meaningful separability when searching for clusters. Other approaches that promote non-normality (non-Gaussianity) include negentropy and mutual information [18] which belong in the broader family of independent component analysis methodologies.

The above approaches nevertheless share the same attribute of offering a single static projection that does not consider any prior knowledge a practitioner may have regarding the high-dimensional latent structure. Such projections can be uninformative as they tend to illustrate the trends of the most evident features which are often already known by the reader. In practise, it has been shown [7] that an interactive or dynamic exploration of the available data can capture better their high-dimensional structure, especially when knowledge from a continual analysis of the data or cohort distribution can be factored into the analysis. Recently, a number of similar methods were introduced to dynamically generate new embeddings guided by the user [32]. Data exploration tasks require studying the patterns and cluster structure of the less dominant features in a dataset. Contrastive PCA (cPCA) [1] computes data projections along which the variance of high-dimensional data is maximized (as in PCA), while the variance of some selected data is minimized. As a result, patterns and variation of the selected data can be removed from the projections. cPCA can be particularly useful when certain subsets of the data convey high variations, which might be obscuring the formation of alternative meaningful patterns in the dataset and thus may hinder its proper exploration.

In this work we focus on computing data projections which filter out any cluster structure associated with some, usually unwanted, prior data knowledge. This knowledge reflects the practitioner's information about the data and by discarding this information from the projections, we hope to uncover unknown patterns. To achieve this, we jointly optimize a projection pursuit index that promotes data separability together with a variance-based objective to remove unwanted background patterns. Our method is applicable in three different prior (background) knowledge case scenarios, where the prior knowledge is represented by a dataset whose structure we wish to remove from the obtained embeddings. Specifically, the cases include:

- **Attribute-based prior**. In this case, the prior data consist of a subset of attributes of the high-dimensional dataset. We want to obtain embeddings that reveal the structure associated with the remaining attributes by discounting the structure of the attributes of the prior dataset. A naive approach to achieve this would be to discard this

---

subset of features from the dataset. However, removing any features from a dataset may not necessarily factor out their variation from the lower dimensional embeddings, as different features are often correlated [20].

- **Sample-based prior**. We wish to visually explore a complex dataset that consists of the combination of a reference dataset (e.g., FMNIST) and a background dataset (e.g., MNIST). The background data modify and/or corrupt the reference dataset and we want to obtain embeddings that remove the background structure and disclose the reference one.
- **Subset-based prior**. The prior dataset consists of a subset of the original high-dimensional samples which are known to be similar to each other. By removing the structure associated with these samples, we can learn embeddings that reveal the structure of the remaining data points. Iterative visual exploration can take place in this setting by repeatedly updating the prior to include the most distinct cluster of each obtained projection.

We specifically propose an efficient and effective optimization on the Stiefel manifold [35], which appears to empirically perform better compared to other related methods and helps circumvent data-related numerical issues. Our main contributions are the following:

- A novel objective function which when optimised computes projections that factor out different types of prior knowledge while also revealing previously unknown underlying structure.
- Manifold optimization modeling of the complex loss function to achieve numerical stability and fast convergence to a desirable solution.
- An iterative framework that can be applied to produce multiple informative projections for the visual data exploration of high-dimensional data.

The rest of the paper is organized as follows. Section 2 expands upon related literature and in Section 3 we give a detailed description of our proposed method. Section 4 provides a quantitative and qualitative analysis of our results and comparison with related approaches. Finally, Section 5 concludes our work.

2 Related Work

A few works on interactive DR have been proposed in the last few years. Contrastive PCA [1] computes data projections which highlight the salient structure of some reference data while discarding the structure of some background data. Conditional t-SNE (ct-SNE) [20] is a generalisation of t-SNE which considers some prior knowledge in order to construct informative 2D embeddings. Prior knowledge corresponds to some information that the user is already aware of and can be represented by a set of labels assigned to the data samples. The labels are either available before any analysis or can be inferred by clustering a set of embeddings. To discount the known factor of the labels, new embeddings are then generated which can provide insight on any underlying unknown structure. In similar fashion, Fair Neighbor Retrieval Visualizer (FAIR-NeRV) [26] computes low-dimensional embeddings which preserve the high-dimensional neighborhood relationships while discarding the reliance on some sensitive information. This information can be specific attributes or class labels and is removed from the embeddings which can then reveal some complementary structure.

Unlike ct-SNE and FAIR-NeRV, which produce non-linear embeddings, SIDE [28] is a linear approach which takes as input some prior knowledge in terms of a background distribution of points.

These points are known to be similar to each other either *a priori* or after some analysis. Projections that promote the maximal difference between the data and the background distribution are then computed. Another recently published method [29] allows the user to guide the examination procedure according to their own exploration interests. The users can formulate their prior knowledge as well as their specific interests in terms of relations among a subset of samples and a subset of attributes. These are then introduced to the model to compute the projections.

Finding lower dimensional embeddings over matrix manifolds has recently become quite popular, mostly due to the flexibility the constraint-free manifold optimization offers [2]. The (compact) Stiefel manifold [35], i.e., the set of all k-tuples of orthonormal vectors, defines a smooth manifold that is endowed with a tangent space with its own inner product, (can be linearized locally around every point) allowing for fast, unconstrained optimization. Matrix manifold optimization has been employed in various cluster exploration applications [38, 17], as well as in general machine learning [8, 34].

3 Methodology

In this section, we first exemplify the optimization details of our method. Secondly, we introduce a framework for structure extraction and knowledge update with which we can perform iterative visual exploration of high-dimensional data.

3.1 *Objective formulation and optimization*

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the column mean-centered data matrix with each row $\{\mathbf{x}_i\}_{i=1}^n$ corresponding to the coordinates of the ith data point. Our goal is to generate a set of low dimensional embeddings $\{\mathbf{q}_i\}_{i=1}^n \in \mathbb{R}^k$ (with $k \ll d$) that render meaningful data separation based on prior information to maximize data cohort informativeness. Kurtosis [9] is a simple and efficient measure of non-normality, which when minimised promotes meaningful separability that tends to reveal informative data cohorts within a dataset. For a set of data projections, the kurtosis is defined as:

$$kurt(\mathbf{X}) = \frac{n \sum_{i=1}^n \left(\mathbf{v}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v} \right)^2}{\left(\mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \right)^2}, \tag{1}$$

where $\mathbf{v} \in \mathbb{R}^d$ is the projection vector. Due to its quartic form, the kurtosis index can have multiple local minima and thus its optimization is a challenging task. Constrained gradient-based methods that force the projection vectors to have unit length [27] have been used to optimize the index but showed poor scalability. An efficient quasi-power method [16, 10, 33] has been designed as an alternative to constrained approaches, which however is less stable in practice when the covariance matrix $\mathbf{X}^\top \mathbf{X}$ is singular. Dimensionality reduction of the original samples (e.g., with SVD) is therefore required so that kurtosis can be optimized on the newly embedded points and can often lead to weak representations. To avoid the singularity issue, we instead optimize kurtosis directly over the Stiefel manifold [35] $St(k, d) \triangleq \{\mathbf{M} \in \mathbb{R}^{d \times k} : \mathbf{M}^\top \mathbf{M} = \mathbf{I}\}$ which ensures unit-length projections over its surface, where an unconstrained optimization can be realised efficiently using any gradient-based solver [2]. We further propose a combination of a kurtosis and a cPCA [1] reconstruction loss. cPCA can effectively maintain data subsets of high-variance while discarding subsets of no interest that may be obscuring the formation of alternative meaningful patterns. Nevertheless, cPCA is not explicitly designed

to provide low-dimensional projections that exhibit meaningful data segregation. As a result, it often happens that cPCA embeddings are not informative. By jointly optimizing kurtosis, improved data separation that can reveal underlying structure is ensured. Let us assume a set of $\{\mathbf{y}_i\}_{i=1}^m$ data samples associated with our prior knowledge, organized in matrix format $\mathbf{Y} \in \mathbb{R}^{m \times d}$. We formulate the above requirements in the following optimization and term our proposed method as IMAPCE (Informative MAnifold Projections for Cluster Exploration).

$$
\mathbf{V}^\star = \operatorname*{arg\,min}_{\mathbf{V} \in \mathbb{R}^{d \times k}} \left\{ f(\mathbf{V}) \triangleq \|\mathbf{X} - \mathbf{X}\mathbf{V}\mathbf{V}^\top\|_F^2 - \alpha \|\mathbf{Y} - \mathbf{Y}\mathbf{V}\mathbf{V}^\top\|_F^2 \right.
$$
$$
\left. + \mu n \sum_{i=1}^n \left[\mathbf{x}_i^\top \mathbf{V}(\mathbf{V}^\top \mathbf{X}^\top \mathbf{X}\mathbf{V})^{-1}\mathbf{V}^\top \mathbf{x}_i \right]^2 \right\}
$$
$$
\text{s.t.} \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}, \tag{2}
$$

where μ is a scaling parameter, α (as in cPCA) regulates the trade-off between having a high target variance and a low background data variance and $\mathbf{V} \in \mathbb{R}^{d \times k}$ with $k < d$ is the projection matrix to be computed. Setting $\alpha = 0$ corresponds to assuming no prior information (and therefore no prior data). $\mathbf{X}, \mathbf{Y}$ are both standardized before the optimization takes place. The gradient of $f(\mathbf{V})$ is given as

$$
\nabla_{\mathbf{V}} f(\mathbf{V}) = 2\left(\alpha\mathbf{Y}^\top\mathbf{Y} - \mathbf{X}^\top\mathbf{X}\right)\mathbf{V}
$$
$$
+ 4\mu n \sum_{i=1}^n \left(\mathbf{x}_i^\top\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^\top\mathbf{x}_i\right)\left[(\mathbf{x}_i\mathbf{x}_i^\top)\mathbf{V}\mathbf{A}^{-1}\right.
$$
$$
\left. - \left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left(\mathbf{V}^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{V}\right)\mathbf{A}^{-1}\right], \tag{3}
$$

where $\mathbf{A} = \mathbf{V}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}$. For visualization purposes we usually set $k = 2, 3$ and in practice therefore $\mathbf{A} \in \mathbb{R}^{k \times k}$ is of small size, rendering its condition number $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ to be relatively small and well-conditioned. By jointly minimizing both objectives of Eq. (2) on the Stiefel manifold, we avoid any singularity issues and at the same time obtain more informative separability of cohorts as we will empirically demonstrate later in the experiments. The time complexity of IMAPCE is $\mathcal{O}(dn^2)$ which mainly owes to the retraction step in the manifold optimization.

3.2 Iterative visual exploration

In the case of a subset-based prior, we are *a priori* aware that a subset of samples of some high-dimensional data are similar to each other (e.g., they share the same class) and we wish to explore the cluster structure of the remaining data points. To achieve this, we optimize Eq. (2), where the kurtosis term is calculated over the remaining (subset of unexplored) samples, defined as $\{\mathbf{z}_i\}_{i=1}^{(n-m)} = \{\mathbf{x}_i\}_{i=1}^n \setminus \{\mathbf{y}_i\}_{i=1}^m$. As a result, the obtained projections provide meaningful data separation of the unexplored samples and reveal their cluster structure.

To extract the structure of the unexplored samples $\mathbf{Z}$, we perform the clustering of their embeddings. While the obtained clusters unveil some previously unknown structure, some of them are not as informative as others. We argue that the cluster which is the most separated from the rest (noted as *most distinct*), is the most informative as it is expected to have the greatest probability of consisting of very similar points. The distinctness of a cluster and its separability are two connected notions stemming from the definition of kurtosis and is naturally encouraged by the optimization. While other cluster

---

Algorithm 1 Iterative Visual Exploration using IMAPCE

Require: original data $\{\mathbf{x}_i\}_{i=1}^n$, prior data $\{\mathbf{y}_i\}_{i=1}^m$, unexplored data $\{\mathbf{z}_i\}_{i=1}^{n-m}$, acceptable cluster size s (for termination criterion).

Ensure: Embeddings $\mathbf{Q}$.

1: **while** $(n - m) > s$ **do**
2:      $\mathbf{V}^\star = \operatorname*{arg\,min}_{\mathbf{V} \in St(k,d)} f(\mathbf{V})$ # IMAPCE
3:      $\mathbf{Q} = \mathbf{Z}\mathbf{V}^\star$ # Embeddings of unexplored data
4:      $\{\mathbf{m}\}_l^r, \{\mathbf{C}\}_l^r \sim DPGMM(\{\mathbf{q}_i\}_{i=1}^{n-m})$ # Clustering
5:      **for** all cluster pairs (l, j) **do**
6:          $\delta_{lj} = \sqrt{(\mathbf{m}_j - \mathbf{m}_l)^\top \mathbf{C}_l^{-1}(\mathbf{m}_j - \mathbf{m}_l)}$
7:          $D_{lj} = (\delta_{lj} + \delta_{jl})/2$
8:      **end for**
9:      $c^\star = \arg\max \mathbf{D1}$ # Most distinct cluster calculation
10:     $\mathbf{X}_c = \{\mathbf{x}_i : \mathbf{x}_i \in c^\star\}$ # Points of the most distinct cluster
11:     $\mathbf{Y} = \mathbf{Y} \cup \mathbf{X}_c$ # Update of prior data and their size m
12:     $\mathbf{Z} = \mathbf{Z}\setminus\mathbf{X}_c$ # Update of unexplored data
13: **end while**

---

characteristics such as shape or density could also be considered, we empirically observed that such characteristics are less reliable across different types of data and features, as opposed to separability.

After finding the most distinct cluster, we dynamically update our prior data $\mathbf{Y}$ to include the points of this cluster, while also removing them from the unexplored subset $\mathbf{Z}$. Optimization of Eq. (2) takes then place to compute new 2D embeddings that exhibit data separation of the updated unexplored data where cluster extraction can again be performed. This process continues iteratively until no more data remain unexplored and can efficiently provide gradual exploration of the unknown underlying structure of high-dimensional data.

To cluster the embeddings we used the Bayesian infinite Gaussian mixture model [30], [4], [24], commonly referred to as Dirichlet Process Gaussian Mixture Model (DPGMM) as it does not require a predefined number of clusters (but rather a maximum cluster number) and also due to its performance quality. More information about the DPGMM is provided in Section A of the Supplementary Material [13].

Given the mean $\mathbf{m}_l$ and covariance matrix $\mathbf{C}_l$ for each cluster l, we can extract the most distinct cluster by calculating all pairwise distances of cluster centers and their respecive distributions using the Mahalanobis distance [22]

$$
\delta_{lj} = \sqrt{(\mathbf{m}_j - \mathbf{m}_l)^\top \mathbf{C}_l^{-1}(\mathbf{m}_j - \mathbf{m}_l)}. \tag{4}
$$

To avoid selecting a small deal of outliers as a cluster, we define a minimum acceptable cluster size. All clusters with size less than that are discarded as outliers and their distances are not considered. From Eq. (4) we define a symmetric pairwise distances matrix for all clusters as $D_{lj} = (\delta_{lj} + \delta_{jl})/2$ and the most distinct cluster is given by

$$
c^\star = \arg\max \mathbf{D1}, \tag{5}
$$

where $\mathbf{1}$ is the vector of all ones. As a special case, if only two clusters of acceptable size are detected, then both are chosen as most distinct. Algorithm 1 outlines the major steps of our proposed iterative framework.

3.3   Hyperparameter tuning

Hyperparamaters α and μ have to be chosen for optimizing Eq. (2). We need to select α if there is any prior data (otherwise it is set to zero), while μ has to be selected regardless of the availabilty of prior data. As a rule of thumb, setting $\alpha = 1$ empirically provides embeddings with a desirable trade-off between high original data variance and low prior data variance. In practice, we observe that the kurtosis term is not greater than 50-100. After computing the cPCA reconstruction error of our original data, μ is selected as one or two orders of magnitude below the cPCA reconstruction error. In this way, kurtosis can influence but not dominate the optimization. To further study the effect of the kurtosis term on the performance of IMAPCE, we perform an ablation study (Section G of the Supplementary Material [13]) on the hyperparameter μ and demonstrate the effectiveness of the proposed empirical rule for choosing μ across different scenarios. Finally, parameter s in iterative visual exploration denotes the minimum acceptable size of a cluster and is used for discarding outliers as well as stopping the exploration process. Its choice depends on the size and dimensionality of the original data.

4   Experimental Results

To showcase that IMAPCE can efficiently factor out different types of prior knowledge, we ran experiments on several datasets for all previously mentioned prior knowledge types. We provide both quantitative and qualitative results and compare IMAPCE's performance against similar embedding methods that promote some underlying structure of the data while also removing any variation associated with the prior knowledge, namely cPCA, ct-SNE and Fair-NeRV.

For cPCA, different projection matrices are computed for a fixed user-selected number of α's (trade-off hyperparameter) and spectral clustering is applied to them. Subsequently, the projection matrix which corresponds to the cluster medoid is used to compute the projections. It is noteworthy that ct-SNE and Fair-NeRV require the input of explicit data labels, which form the prior information to be removed from the generated embeddings. Details about the hyperparameters' selection for ct-SNE and Fair-NeRV are provided in Section B of the Supplementary Material [13].

We implemented IMAPCE on Python and used Pymanopt [39], a Python toolbox for optimization on Riemannian manifolds. The DPGMM used for clustering is implemented via the Sci-kit learn library of Python [25]. Our implementation is provided in https://github.com/StavGer/IMAPCE. For cPCA, ct-SNE and Fair-NeRV we used the official implementations. Experimental results are organized according to different prior knowledge cases.

4.1   Attribute-based prior

The task in this case is to generate embeddings that factor out the structure of one or more selected attributes of the high-dimensional data and reveal the structure of the remaining attributes. We ran experiments using IMAPCE, ct-SNE, cPCA and Fair-NeRV and compare their performance on both synthetic and real-world data.

4.1.1   Synthetic data

The synthetic dataset [14] consists of 1,500 ten-dimensional points. All points are assigned to one of two clusters (with centers sampled from $\mathcal{N}(0, 25)$) in the first four dimensions and one of three clusters (with centers sampled from $\mathcal{N}(0, 1)$) in dimensions five and six.

For each point we add noise from $\mathcal{N}(0, 0.01)$. The last four dimensions correspond to samples from $\mathcal{N}(0, 1)$. To run IMAPCE (we set $\alpha = 1$, $\mu = 200$) and cPCA, we define as prior data the first four dimensions and wish to generate embeddings that reveal the complementary cluster structure of dimensions five and six. With the same goal, we implement ct-SNE and Fair-NeRV using the cluster labels of the first four dimensions as prior.

The generated embeddings are shown in figures 1(a)-1(d) where prior data information is encoded according to shape and complementary structure according to color. We observe that cPCA embeddings are clustered with respect to their shape, indicating that the structure of prior data is not removed. On the contrary, ct-SNE, Fair-NeRV and IMAPCE compute embeddings that factor out the prior information as there is mix of points with different shapes. Fair-NeRV and IMAPCE seem to achieve a slightly better separation with respect to the green and red cluster than ct-SNE, unveiling the complementary structure to a better extent.

The Normalised Laplacian Score (NLS) was proposed [20] to quantify the presence of some prior label information on a set of embeddings. This score takes values in [0, 1] and measures the label homogeneity within a user-selected neighborhood in an embedding set. Higher NLS values indicate better prior information removal w.r.t some labels from the embeddings. To quantitatively compare how well the embeddings of all methods factor out the labels of dimensions one to four, we calculated the NLS with respect to these labels and for neighborhood sizes of 10, 20, . . . , 100. We report the average NLS over the neighborhood size in Table 1. IMAPCE achieved higher mean NLS (lower homogeneity) than cPCA and ct-SNE, indicating that it removes the prior information more effectively (as also observed qualitatively). We argue that the inclusion of the kurtosis term in IMAPCE achieves the meaningful data separation that reveals some unknown structure missed by cPCA. Fair-NeRV achieves a slightly bigger average NLS than IMAPCE, nevertheless at a higher computational cost as shown later in Section 4.2.

4.1.2   UCI Adult data

We sampled 1,000 data points from the UCI Adult dataset [5] which consists of six features. Age, education level and work hours per week are numeric while ethnicity (white/other), gender (male/female) and income ($> 50,000$) are binary.

Using the ethnicity feature as prior, we obtain cPCA, ct-SNE, Fair-NeRV and IMAPCE ($\alpha = 1, \mu = 150$) embeddings as shown in Figures 1(e) - 1(h). cPCA provides embeddings with mixed ethnicity, gender and income features, failing to exhibit any clear cluster formation. ct-SNE fails to remove the ethnicity information from male instances, but successfully does so for female ones. In addition, the complementary gender structure is not unveiled as there is overlap between red and green points. On the contrary, Fair-NeRV and IMAPCE remove the ethnicity variation by having clusters with mixed ethnicities. Their results are similar as they exhibit clusters of different gender (red-green colors) and income (filled-unfilled markers) attributes. However, IMAPCE slightly outperforms Fair-NeRV which incorrectly clusters together points with different gender and some points with different income labels.

Similar to the synthetic data experiments, we computed the NLS scores of all methods with respect to the ethnicity attribute and for neighborhood sizes of 10, 20, . . . , 100. We report the average NLS over the neighborhood size in Table 1. IMAPCE and Fair-NeRV achieve higher average NLS and thus discount the ethnicity attribute information more effectively than both cPCA and ct-SNE, as also

qualitatively confirmed. Similar experiments using gender attribute and the combination of gender and ethnicity attributes as prior are included in Section C of the Supplementary Material [13].

Table 1: Mean test accuracy of a Linear classifier over ten different random train-test splits for both types of Complex data and mean NLS over 10, 20, ..., 100 neighborhood sizes using the specified attribute(s) as prior for synthetic and adult.

Dataset	cPCA	ct-SNE	Fair-NeRV	IMAPCE
Synthetic	0.16	0.34	**0.49**	0.45
Adult	0.11	0.13	**0.23**	0.21
MNIST+FMNIST	0.75	0.65	0.78	**0.81**
CIFAR 100+FMNIST	0.58	0.79	0.83	**0.95**

Table 2: Execution time of all methods with respect to each dataset.

Dataset	cPCA	ct-SNE	Fair-NeRV	IMAPCE
Synthetic	**2 sec**	30 sec	5 min	3 sec
Adult	**5 sec**	70 sec	5 min	**5 sec**
MNIST+FMNIST	**21 sec**	63 sec	1hr-cutoff	26 sec
CIFAR+FMNIST	**22 sec**	61 sec	1hr-cutoff	25 sec

4.2   Sample-based prior

4.2.1   MNIST + Fashion-MNIST (FMNIST)

We created some complex data by combining instances from MNIST (which we consider the background part) and FMNIST (which we consider as the reference part) datasets. The task in this case is to compute two-dimensional embeddings that remove the information associated with the MNIST data and provide separation according to the complementary structure defined by the FMNIST labels. We constructed a series of complex data (6,000 samples) as follows. We first selected two different FMNIST classes and randomly sampled 6,000 images of those classes. We then superimposed those images with randomly chosen MNIST images. We repeated this procedure for a few combinations of FMNIST classes in order to create different series of complex data. An example of a complex data instance and its construction is shown in Figure 2(a). The superimposed 28x28 images, as well as the MNIST images serving as the background, are flattened to 784 dimensional vectors before their processing. As background data we randomly select 1,000 MNIST samples which are different from the ones used for the complex data formation, but adequately capture the MNIST variation we wish to remove from the embeddings. While we can employ this prior data for both IMAPCE and cPCA, ct-SNE and Fair-NeRV are limited to label priors. To remove the MNIST information using ct-SNE and Fair-NeRV, we utilize the ground truth labels of the MNIST instances that were used for the complex data construction. Finally, for IMAPCE we set $\alpha = 1$ and $\mu = 10^5$.

Figures 2(b)-2(e) show complex MNIST + FMNIST data embeddings (with "Sandal", "Sneaker" FMNIST ground truth labels) generated by all methods. We observe that ct-SNE, cPCA and Fair-NeRV separate a small portion of "Sandal" embeddings from the "Sneaker" ones. However, there is significant overlap between the majority of

"Sandal" and "Sneaker" instances. On the other hand, IMAPCE computes embeddings that provide the clearest separation with respect to their FMNIST class as their overlap is smaller than all other methods.

To quantify the separability according to the FMNIST labels, we trained (75%) and tested (25%) a Linear classifier on the classification of the calculated 2D complex data embeddings with respect to these labels. The test-set accuracy is averaged over 10 random train-test splits and given in Table 1. IMAPCE not only achieves the best mean test-set accuracy and thus separation (also observed visually), but at the same time considers only a small amount of background MNIST information, as opposed to ct-SNE and Fair-NeRV which require all MNIST labels. We provide additional experiments for different MNIST + FMNIST complex data as well as SVM classifiers with accuracy, F1 and Recall scores in Section D of the Supplementary Material [13].
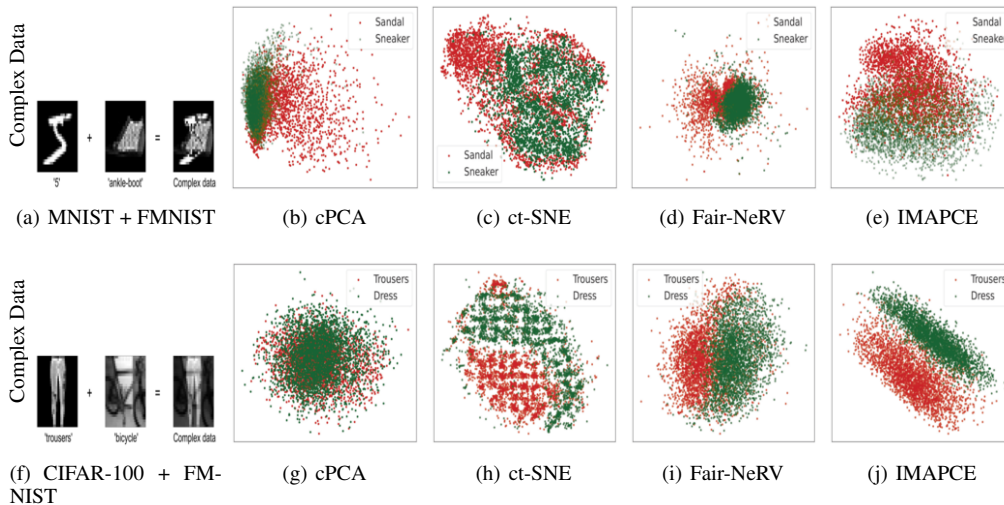
4.2.2   CIFAR 100 + FMNIST

Similar to the combination of MNIST with FMNIST data, we constructed more complex data by superimposing 6,000 images from CIFAR-100 [21] with images from FMNIST. CIFAR-100 has 100 classes, each of which has 600 color images. To perform the superimposition of 32x32 CIFAR-100 images with the 28x28 FMNIST images, we cropped the border of the CIFAR-100 images and subsequently converted them from color (RGB) to grayscale. Afterwards, we flattened each 28x28 superimposed image to a 784 dimensional vector. An example of a grayscale CIFAR-100 instance, FMNIST instance and their superimposition is shown in Figure 2(f), while more examples are provided in Section E of the Supplementary Material [13]. In this case, we consider the CIFAR-100 image data as background and the FMNIST data as the reference. Therefore, we wish to compute embeddings which exhibit cluster formation according to their FMNIST label. To achieve this, the CIFAR-100 label variation should be removed from the generated embeddings. For cPCA and IMAPCE, we define as background data 5,000 randomly selected CIFAR-100 images (not necessarily used during the dataset construction) which can capture the variance of the CIFAR-100 data. On the contrary, for ct-SNE and Fair-NeRV we provide the CIFAR-100 ground truth labels of the CIFAR-100 images we used when constructing the complex data. For IMAPCE, we set $\alpha = 1$ and $\mu = 10^7$.

Figures 2(g)-2(j) show complex CIFAR-100 + FMNIST data embeddings (with "Trousers", "Dress" FMNIST ground truth labels) generated by all methods. We observe that cPCA completely fails to separate the embeddings according to their FMNIST class. While ct-SNE and Fair-NeRV achieve some sort of separation, there is considerable overlap of "Trousers" and "Dress" instances. On the contrary, the embeddings computed by IMAPCE achieve the best separation regarding their FMNIST class.

To complement our qualitative observations, we trained a Linear classifier on the classification of the calculated 2D embeddings of all methods using identical specifications to the classifier of the MNIST + FMNIST case. The test-set accuracies are provided in Table 1, where the superiority of IMAPCE among all methods is demonstrated. This superiority is boosted by the fact that it is achieved under an unsupervised setting by considering only a small number (5,000 out of the total 60,000) of representative CIFAR-100 background instances that can effectively capture the CIFAR-100 structure that has to be factored out from the computed embeddings. On the contrary, both ct-SNE and Fair-NeRV need the specific CIFAR-100 labels of all the instances that were used when constructing the complex data.

(a) cPCA   (b) ct-SNE   (c) Fair-NeRV   (d) IMAPCE

(e) cPCA   (f) ct-SNE   (g) Fair-NeRV   (h) IMAPCE
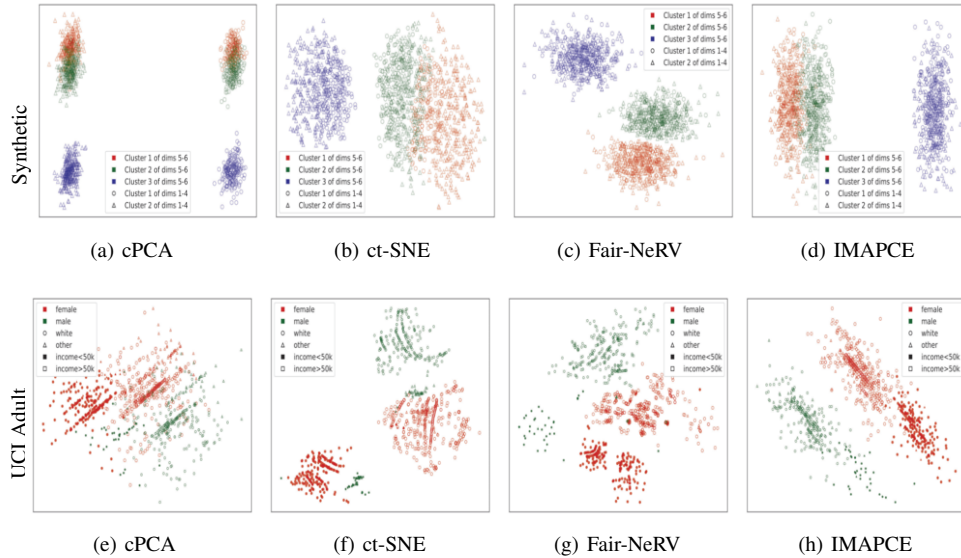
Figure 1: Top row shows synthetic data experiments with information of dimensions one to four as a prior. Bottom row illustrates UCI adult data experiments with ethnicity feature as a prior. (a) cPCA embeddings clustered w.r.t. labels from dimensions one to four. (b) ct-SNE embeddings are clustered w.r.t. labels from dimensions five to six (complementary structure) with some noticeable error (overlap). (c,d) Fair-NeRV and IMAPCE embeddings are clustered w.r.t. labels from dimensions five to six (complementary structure) with clearer separation. (e) cPCA fails to separate embeddings according to their gender and income. (f) ct-SNE clearly separates embeddings w.r.t. to their income and to some extent according to their gender. (g) Fair-NeRV computes embeddings that are mostly separated w.r.t. to gender and income with the exception of some outliers, while (h) IMAPCE perfectly clusters embeddings according to their gender and income (revealing complementary structure).


(a) MNIST + FMNIST   (b) cPCA   (c) ct-SNE   (d) Fair-NeRV   (e) IMAPCE

(f) CIFAR-100 + FM-NIST   (g) cPCA   (h) ct-SNE   (i) Fair-NeRV   (j) IMAPCE

Figure 2: Top row shows complex MNIST + FMNIST experiments using MNIST as prior (a), while bottom row complex CIFAR-100 + FMNIST experiments using CIFAR-100 as prior (f). In both cases, embeddings by cPCA (b,g), ct-SNE (c,h) and Fair-Nerve (d,i) appear significantly more mixed w.r.t their class, as opposed to the IMAPCE ones (e,j) which exhibit better segregation.

Finally, in terms of running time performance, IMAPCE has a comparable execution time to cPCA, while it is faster than both ct-SNE and Fair-NeRV across all experimental setups as Table 2 shows. Due to the large feature space of the sample-based prior experiments, an 1hr cutoff time was set for Fair-NeRV (and the final embeddings computed within this cutoff time were used for the experiments) to facilitate hyperparameter search. More experiments for different combinations of FMNIST classes are included in Section E of the Supplementary Material [13]. Additionally, SVM classifiers are trained on the 2D embeddings and accuracy, F1 and Recall scores are computed for all methods and are also included in the same section.

4.3 Subset-based prior

In this setup, prior data are subsets of the original data that share the same class (and are thus similar). The goal is to generate embeddings that reveal the cluster structure of the remaining samples (unexplored subset) of these datasets. By employing our structure extraction framework with IMAPCE or cPCA, we sequentially obtain new sets of informative embeddings that gradually unveil the underlying cluster structure of all high-dimensional data. In this case, we did not compare against ct-SNE and Fair-NeRV as they both require data label availability for the whole dataset, which contradicts the present subset-based analysis setup.

We performed the iterative visual exploration of Image Segmentation data from the UCI machine learning repository [11] under vari-
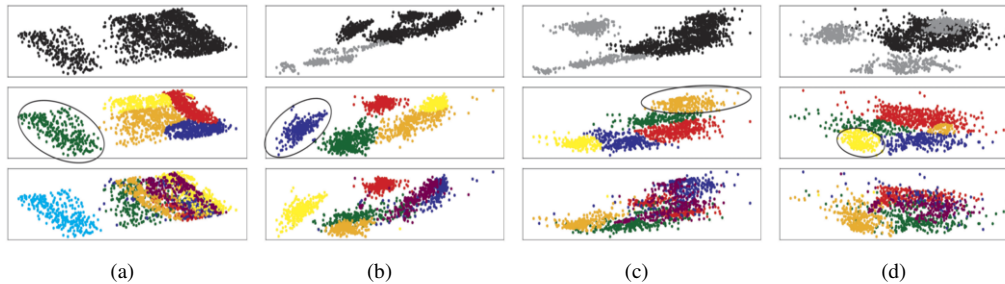
Figure 3: Iterative exploration of UCI image segmentation data by IMAPCE. Each column (subfigure) corresponds to an iteration of the exploration process. In every column (subfigure), the upper plot shows the data projections where prior data are colored in grey while the unexplored subset in black. The middle plot shows the clustering of the unexplored points according to DPGMM and the most distinct cluster encircled. The bottom plot illustrates the unexplored points colored according to their ground truth label.

ous prior data assumptions. This dataset consists of 2,310 samples and 19 attributes and includes 330 instances from seven different classes, namely "sky", "grass", "path", "foliage", "cement", "brick-face", "window". Assuming no prior data and selecting $\alpha = 1$, $\mu = 10^5$, $s = 75$, we sequentially apply IMAPCE on UCI Image segmentation data. The obtained projections are shown in Figure 3, where each subfigure consists of three subplots. The top row illustrates the IMAPCE data projections for a specific iteration of the process. Grey points correspond to the prior data, while black points correspond to the unexplored subset of samples. The middle row shows the results of a DPGMM clustering on the unexplored points where the most distinct cluster is encircled. The bottom row shows the unexplored points coloured according to their ground truth class.

The first data projection is shown in Figure 3(a). The upper plot consists of solely black points as there is no prior data. All data are clustered in the middle subplot and the green cluster is the most distinct. Its points are considered very similar and are stored for the evaluation stage. Subsequently, these points define the prior data and are removed from the unexplored data. Afterwards, the second iteration takes place and new informative embeddings are calculated and shown in Figure 3(b). Separation of data samples that were previously overlapping is now observed and indicative of an informative data projection. The grey points of the upper subplot correspond to the prior data while the black ones refer to the unexplored points, which are clustered in the middle subplot. The most distinct cluster is the blue one and its points are incorporated in the prior data and removed from the unexplored data. This iterative process is repeated until no clusters of the unexplored points are formed (only outliers are left). Gradual exploration of the whole dataset contributes to the extraction of new and meaningful underlying structure. For brevity and demonstration reasons, only the first four iterations are illustrated in Figures 3(a)-3(d) and the rest of the exploration analysis is omitted.

To quantitatively compare IMAPCE and cPCA, we ran several experiments on UCI Image Segmentation ($\alpha = 1$, $\mu = 10^5$, $s = 75$) with different initial subsets of prior data. Performance evaluation took place after the exploration of the data had finished. During the evaluation stage, the quality of the most distinct clusters (which are stored along the exploration process) is measured with respect to their ground truth labels using the Jaccard [19] and NMI scores. Both scores are widely used in the DR literature for evaluating cluster quality. Detailed results for IMAPCE and cPCA are given in Table 3. We consider the case of no prior data and also experiment by setting as prior data subsets of data samples from every ground truth class. Overall, IMAPCE clearly outperforms cPCA on both scores under all

prior data assumptions. The superiority of IMAPCE on the Jaccard and NMI scores indicates that it promotes enhanced cluster segregation in comparison to cPCA, owing to the inclusion of the kurtosis term. We provide equivalent quantitative results for 10,000 randomly selected MNIST instances in Section F of the Supplementary Material [13].

Table 3: Jaccard (averaged over all classes) and NMI scores for the evaluation of IMAPCE and cPCA on the visual exploration of the UCI Image Segmentation data under different prior data assumptions. Ten random initialisations were used for IMAPCE.

Method	Prior class	Mean Jaccard	NMI
cPCA	-	0.44	0.46
IMAPCE	-	**0.63** ± 0.05	**0.67** ± 0.06
cPCA	"Brickface"	0.49	0.60
IMAPCE	"Brickface"	**0.63** ± 0.04	**0.67** ± 0.06
cPCA	"Cement"	0.47	0.60
IMAPCE	"Cement"	**0.64** ± 0.07	**0.69** ± 0.06
cPCA	"Foliage"	0.47	0.61
IMAPCE	"Foliage"	**0.65** ± 0.05	**0.72** ± 0.05
cPCA	"Grass"	0.50	0.48
IMAPCE	"Grass"	**0.59** ± 0.08	**0.64** ± 0.08
cPCA	"Window"	0.50	0.57
IMAPCE	"Window"	**0.68** ± 0.06	**0.74** ± 0.04
cPCA	"Sky"	0.35	0.43
IMAPCE	"Sky"	**0.57** ± 0.05	**0.61** ± 0.07
cPCA	"Path"	0.43	0.56
IMAPCE	"Path"	**0.59** ± 0.06	**0.65** ± 0.07

5 Conclusion

In this work we proposed IMAPCE, a dimensionality reduction method to generate low-dimensional embeddings that reveal previously unknown underlying structures while at the same time filtering out any unwanted prior information from the original data. A novel bi-objective formulation is introduced that jointly optimizes a projection pursuit index with a contrastive-PCA loss over the Stiefel manifold. We validated the performance of our proposed method on a set of diverse datasets for three different case scenarios of prior knowledge removal. IMAPCE demonstrated competitive performance against other compared methods in both cluster separation and removal of unwanted knowledge without requiring explicit data labels, while also being computationally efficient in practice.

Acknowledgements

References

[1] A. Abid, M. J. Zhang, V. K. Bagaria, and J. Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, May 2018.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2007.

[3] E. Amid and M. K. Warmuth. TriMap: Large-scale Dimensionality Reduction Using Triplets. *arXiv preprint arXiv:1910.00204*, 2019.

[4] J. R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429, 1991.

[5] B. Becker and R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[7] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, 2019.

[8] H. E. Cetingul and R. Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1896–1902. IEEE, 2009.

[9] S.-S. Chiang, C.-I. Chang, and I. Ginsberg. Unsupervised target detection in hyperspectral images using projection pursuit. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7):1380–1391, 2001.

[10] S. Driscoll, Y. MacMillan, and P. Wentzell. Sparse projection pursuit analysis: An alternative for exploring multivariate chemical data. *Analytical Chemistry*, 12 2019.

[11] D. Dua and C. Graff. UCI machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

[12] J. Friedman and J. Tukey. A projection pursuit algorithm for exploratory data analysis," ieee transactions on computers, c-23, 881-889. *Computers, IEEE Transactions on*, C 23:881 – 890, 10 1974.

[13] S. Gerolymatos, X. Evangelopoulos, V. Gusev, and J. Y. Goulermas. Cluster exploration using informative manifold projections, 2024. URL https://arxiv.org/abs/2309.14857.

[14] E. Heiter, B. Kang, R. Seurinck, and J. Lijffijt. Revised conditional t-sne: Looking beyond the nearest neighbors. In *Advances in Intelligent Data Analysis XXI*, page TBD, Cham, 2023. Springer International Publishing.

[15] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–520, 1933.

[16] S. Hou and P. Wentzell. Fast and simple methods for the optimization of kurtosis used as a projection pursuit index. *Analytica chimica acta*, 704:1–15, 2011.

[17] L.-K. Huang and S. Pan. Communication-efficient distributed PCA by Riemannian optimization. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4465–4474. PMLR, 13–18 Jul 2020.

[18] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, 13 4-5:411–30, 2000.

[19] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:241–72, 01 1901.

[20] B. Kang, D. G. García, J. Lijffijt, R. Santos-Rodríguez, and T. de Bie. Conditional t-sne: More informative t-sne embeddings. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–2, 2021.

[21] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[22] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[23] L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018.

[24] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249 – 265, 2000.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] J. Peltonen, W. Xu, T. Nummenmaa, and J. Nummenmaa. Fair neighbor embedding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[27] D. Peña and F. J. Prieto. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3):286–300, 2001.

[28] K. Puolamäki, E. Oikarinen, B. Kang, J. Lijffijt, and T. De Bie. Interactive visual data exploration with subjective feedback: An information-theoretic approach. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1208–1211, 2018.

[29] K. Puolamäki, E. Oikarinen, and A. Henelius. Guided visual exploration of relations in data sets. *Journal of Machine Learning Research*, 22(96): 1–32, 2021.

[30] C. E. Rasmussen. The infinite gaussian mixture model. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 554–560. MIT Press, 1999.

[31] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science (New York, N.Y.)*, 290:2323–6, 2001.

[32] D. A. Senanayake, W. Wang, S. H. Naik, and S. K. Halgamuge. Self-organizing nebulous growths for robust and incremental data visualization. *IEEE Transactions on Neural Networks and Learning Systems*, 32: 4588–4602, 2019.

[33] H. Siyuan and P. Wentzell. Re-centered kurtosis as a projection pursuit index for multivariate data analysis. *Journal of Chemometrics*, 28, 2014.

[34] A. Smith, B. Laubach, I. Castillo, and V. M. Zavala. Data analysis using riemannian geometry and applications to chemical engineering. *Computers & Chemical Engineering*, 168:108023, 2022.

[35] E. Stiefel. Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten. *Commentarii mathematici Helvetici*, 8:305–353, 1935.

[36] J. Tang, J. Liu, M. Zhang, and Q. Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, apr 2016.

[37] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319, 2000.

[38] F. J. Theis, T. P. Cason, and P. A. Absil. Soft dimension reduction for ica by joint diagonalization on the stiefel manifold. In T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, editors, *Independent Component Analysis and Signal Separation*, pages 354–361. Springer Berlin Heidelberg, 2009.

[39] J. Townsend, N. Koep, and S. Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

[40] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.