Non-redundant Spectral Dimensionality Reduction

Yochai Blau[(⊠)] and Tomer Michaeli

Technion–Israel Institute of Technology, Haifa, Israel
`yochai@campus.technion.ac.il`, `tomer.m@ee.technion.ac.il`

Abstract. Spectral dimensionality reduction algorithms are widely used in numerous domains, including for recognition, segmentation, tracking and visualization. However, despite their popularity, these algorithms suffer from a major limitation known as the "repeated eigen-directions" phenomenon. That is, many of the embedding coordinates they produce typically capture the same direction along the data manifold. This leads to redundant and inefficient representations that do not reveal the true intrinsic dimensionality of the data. In this paper, we propose a general method for avoiding redundancy in spectral algorithms. Our approach relies on replacing the orthogonality constraints underlying those methods by unpredictability constraints. Specifically, we require that each embedding coordinate be unpredictable (in the statistical sense) from all previous ones. We prove that these constraints necessarily prevent redundancy, and provide a simple technique to incorporate them into existing methods. As we illustrate on challenging high-dimensional scenarios, our approach produces significantly more informative and compact representations, which improve visualization and classification tasks.

1 Introduction

The goal in nonlinear dimensionality reduction is to construct compact representations of high dimensional data, which preserve as much of the variability in the data as possible. Such techniques play a key role in diverse applications, including recognition and classification [3,12,18], tracking [24,25,38], image and video segmentation [21,28], pose estimation [11,29], age estimation [15], spatial and temporal super-resolution [7,28], medical image and video analysis [5,34] and data visualization [26,37,40].

Many of the dimensionality reduction methods developed in the last two decades are based on spectral decomposition of some data-dependent (kernel) matrix. These include, e.g., Locally Linear Embedding (LLE) [30], Laplacian Eigenmaps (LEM) [2], Isomap [35], Hessian Eigenmaps (HLLE) [9], Local Tangent Space Alignment (LTSA) [41], Diffusion Maps (DFM) [8], and Kernel
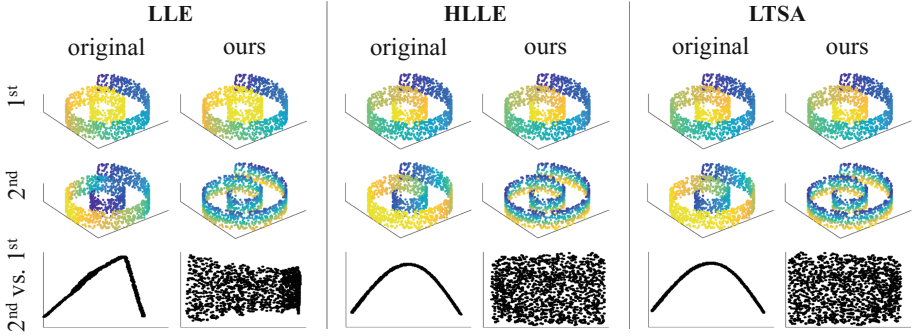

Fig. 1. The first two projections of data points lying on a Swiss roll manifold, as obtained with the original LLE, HLLE and LTSA algorithms and with our non-redundant versions of those algorithms. Top rows: the points colored by the projections. The original algorithms redundantly capture progression along the angular direction twice. In contrast, with our modifications, the second projection captures the vertical direction. Bottom row: scatter plot of the 2nd projection vs. the 1st. In the original algorithms, the 2nd projection is a function of the 1st, while in our algorithms it is not. (Color figure online)

Principal Component Analysis (KPCA) [32]. Methods in this family differ in how they construct the kernel matrix, but in all of them the eigenvectors of the kernel serve as the low-dimensional embedding of the data points [4,17,36].

A significant shortcoming of spectral dimensionality reduction algorithms is the "repeated eigen-directions" phenomenon [10,13,14]. That is, successive eigenvectors tend to represent directions along the data manifold which were already captured by previous ones. This leads to redundant representations that are unnecessarily larger than the intrinsic dimensionality of the data. To illustrate this effect, Fig. 1 visualizes the two dimensional embeddings of a Swiss roll, as obtained by several popular algorithms. In all the examined methods, the second dimension of the embedding carries no additional information with respect to the first. Specifically, although the first dimension already completely characterizes the position along the long axis (angular direction) of the manifold, the second dimension is also a function of this axis. Progression along the short axis (vertical direction) is captured only by the third eigenvector in this case (not shown). Therefore, the two dimensional representation we obtain is 50% redundant: Its second feature is a deterministic function of the first.

In fact, the redundancy of spectral methods can be arbitrarily high. To see this, consider for example the embedding obtained by the LEM method, whose kernel approximates the Laplace-Beltrami operator on the manifold. The Swiss-roll corresponds to a two dimensional strip with edge lengths L_1 and L_2. Thus, the eigenfunctions and eigenvalues (with Neumann boundary conditions) are given in this case by
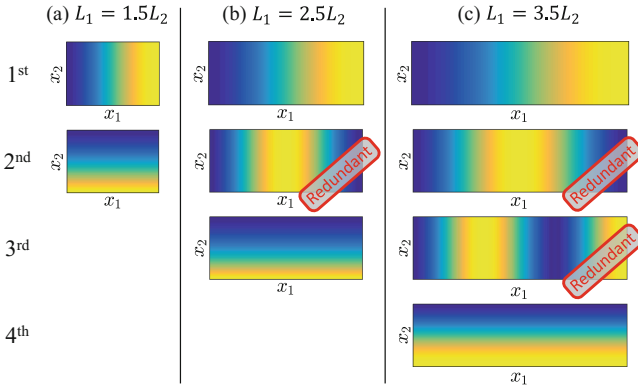
Fig. 2. A 2D strip with edge lengths (a) $L_1 = 1.5L_2$, (b) $L_1 = 2.5L_2$ and (c) $L_1 = 3.5L_2$, colored according to the first few coordinates of the Laplacian Eigenmaps embedding. Coordinates $2, \ldots, \lfloor L_1/L_2 \rfloor$ are redundant as they are all functions of only x_1, which is already fully represented by the first coordinate. (Color figure online)

$$\phi_{k_1 k_2}(x_1, x_2) = \cos\left(\frac{k_1 \pi x_1}{L_1}\right) \cos\left(\frac{k_2 \pi x_2}{L_2}\right), \tag{1}$$

$$\lambda_{k_1 k_2} = \left(\frac{k_1 \pi}{L_1}\right)^2 + \left(\frac{k_2 \pi}{L_2}\right)^2, \tag{2}$$

for $k_1, k_2 = 0, 1, 2, \ldots$, where x_1 and x_2 are the coordinates along the strip. Ignoring the trivial function $\phi_{0,0}(x_1, x_2) = 1$, it can be seen that the first $\lfloor L_1/L_2 \rfloor$ eigenfunctions (corresponding to the smallest eigenvalues) are functions of only x_1 and not x_2 (see Fig. 2). Thus, at least $\lfloor L_1/L_2 \rfloor + 1$ projections are required to capture the two dimensions of the manifold, which leads to a very inefficient representation when L_1 is much larger than L_2. Projections $2, \ldots, \lfloor L_1/L_2 \rfloor$ are all functions of projection 1, and are thus redundant. For example, when $L_1 > 2L_2$, the first two eigenfunctions are $\phi_{1,0}(x_1, x_2) = \cos(\pi x_1/L_1)$ and $\phi_{2,0}(x_1, x_2) = \cos(2\pi x_1/L_1)$, which clearly satisfy $\phi_{2,0}(x_1, x_2) = 2\phi_{1,0}^2(x_1, x_2) - 1$. Notice that this redundancy appears despite the fact that the functions $\{\phi_{k_1 k_2}\}$ are orthogonal. This highlights the fact that *orthogonality does not imply non-redundancy*.

The above analysis is not unique to the LEM method. Indeed, as shown in [14], spectral methods produce redundant representations whenever the variances of the data points along different manifold directions vary significantly. This observation, however, cannot serve to solve the problem as in most cases the underlying manifold is not known a-priori.

In this paper, we propose a general framework for eliminating the redundancy caused by repeated eigen-directions. Our approach applies to all spectral dimensionality reduction algorithms, and is based on replacing the orthogonality constraints underlying those methods, by unpredictability ones. Namely, we restrict subsequent projections to be unpredictable (in the statistical sense) from all previous ones. As we show, these constraints guarantee that the projections
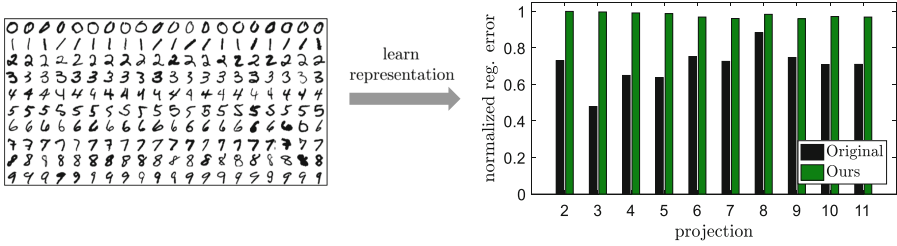
Fig. 3. A 10-dimensional representation of 15K MNIST handwritten digits [23] was learned with LEM and our non-redundant LEM. The bar plots show the normalized errors attained in regressing each projection against all previous ones, indicating to what extent the projection is redundant (higher is less redundant) [10].

be non-redundant. Therefore, once a manifold dimension is fully represented by a set of projections in our method, the following projections must capture a new direction along the manifold. As we demonstrate on several high-dimensional data-sets, the embeddings produced by our algorithm are significantly more informative than those learned by conventional spectral methods.

2 Related Work

Very few works suggested ways to battle the repeated eigen-directions phenomenon. Perhaps the simplest approach is to identify the redundant projections in a post-processing manner [10]. In this method, one begins by computing a large set of projections. Each projection is then regressed against all previous ones (via nonparametric regression). Projections with low regression errors (i.e. which can be accurately predicted from the preceding ones) are discarded. This approach is quite efficient but usually works well only in simple situations. Its key limitation is that it is restricted to choose the projections from a given finite set of functions, which may not necessarily contain a "good" subset. Indeed, as we demonstrate in Fig. 3, in real-world high-dimensional settings all the projections tend to be partially predictable from previous ones. Yet, there usually does not exist any single projection which can be considered fully redundant. Therefore, despite the obvious dependencies, almost no projection is practically discarded in this approach. In contrast, our algorithm produces projections which cannot be predicted from the previous ones (with normalized regression errors ∼100%). Therefore, we are able to preserve more information about the data.

Another simple approach is to compute the projections sequentially, by eliminating the variations in the data which can be attributed to the projections that have already been computed. A naive way of doing so, would be to subtract from the data points their reconstructions based on all the previous projections. However, perhaps counter-intuitively, this *sequential regression* process does not necessarily prevent redundancy. This is because the data points may fall off the manifold during the iterations, as demonstrated in Fig. 4(b).
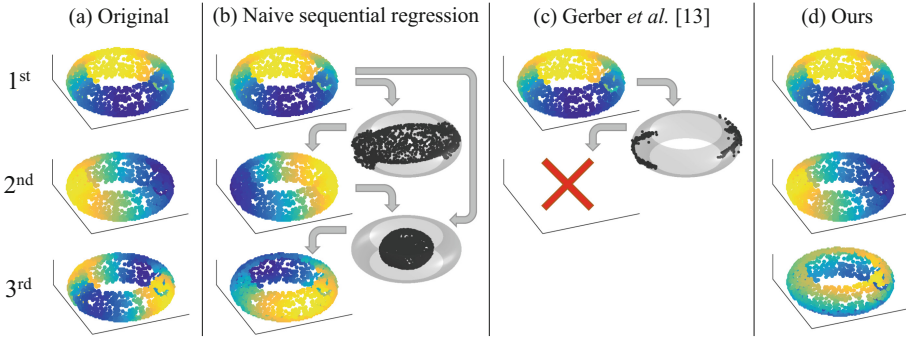
Fig. 4. (a) The first three projections of points lying on a ring manifold, obtained with the original LEM algorithm. The projections correspond to $\cos(\theta)$, $\sin(\theta)$ and $\sin(2\theta + c)$, where θ is the outer angle of the ring. In this case, Projection 2 is not a function of Projection 1 and is thus non-redundant. But Projection 3 is a function of Projections 1 and 2, and is thus redundant. (b) The projections obtained with the naive sequential regression approach (Sect. 2). Here, Projection 3 is still redundant. The right column shows the points after subtracting their prediction from previous projections, which causes them to fall off the manifold. (c) The projections obtained with the algorithm of [13]. Here, the algorithm halts after one projection. The right column shows the points after the advection process along the manifold, which results in two clusters forming an unconnected graph. (d) The projections obtained with our non-redundant version of LEM. Our algorithm extracts a non-redundant third projection, which captures progression along the inner angle of the ring.

A more sophisticated approach, suggested by Gerber et al. [13], is to collapse the data points *along the manifold* in the direction of the gradient of the previous projection. In this approach, the points always remain on the manifold. However, this method fails whenever a projection is a non-monotonic function of some coordinate along the manifold. This happens, for example, in the ring manifold of Fig. 4. In this case, the first projection extracted by LEM corresponds to $\cos(\theta)$, where θ is the outer angle of the ring. Therefore, before computing the second projection, the advection process moves the points along the θ coordinate towards the locations at which $\cos(\theta)$ attains its mean value, which is 0. This causes the points with $\theta \in (0, \pi)$ to collapse to $\theta = \pi/2$, and the points with $\theta \in (\pi, 2\pi)$ to collapse to $\theta = 3\pi/2$. The two resulting clusters form an unconnected graph, so that LEM cannot be applied once more. An additional drawback of this method is that it requires a-priori knowledge of the manifold dimension. Furthermore, it is very computationally intensive and thus impractical for high-dimensional big data applications.

In this paper, we propose a different approach. Similarly to the methods described above, our algorithm is sequential. However, rather than heuristically modifying the data points in each stage, we propose to directly incorporate constraints which guarantee that the projections are not redundant.

3 Eliminating Redundancy

Nonlinear dimensionality reduction algorithms seek a set of *non-linear* projections $f_i : \mathbb{R}^D \to \mathbb{R}$, $i = 1, \cdots, d$ which map D-dimensional data points $\boldsymbol{x}_n \in \mathbb{R}^D$ into a d-dimensional feature space ($d < D$).

Definition 1. *We call a sequence of projections $\{f_i\}$ **non-redundant** if none of them can be expressed as a function of the preceding ones. That is, for every i,*

$$f_i(\boldsymbol{x}) \neq g(f_{i-1}(\boldsymbol{x}), \cdots, f_1(\boldsymbol{x})) \tag{3}$$

for every function $g : \mathbb{R}^{i-1} \to \mathbb{R}$.

Let us see why existing spectral dimensionality reduction algorithms do not necessarily yield non-redundant projections. Spectral algorithms obtain the ith projection of all the data points, denoted by $\boldsymbol{f}_i = (f_i(\boldsymbol{x}_1), \cdots, f_i(\boldsymbol{x}_N))^T$, as the solution to the optimization problem[1]

$$\begin{aligned} \max_{\boldsymbol{f}_i} \quad & \boldsymbol{f}_i^T \boldsymbol{K} \boldsymbol{f}_i \\ \text{s.t.} \quad & \boldsymbol{1}^T \boldsymbol{f}_i = 0 \\ & \boldsymbol{f}_i^T \boldsymbol{f}_i = 1 \\ & \boldsymbol{f}_i^T \boldsymbol{f}_j = 0, \quad \forall j < i. \end{aligned} \tag{4}$$

Here, $\boldsymbol{K}$ is an $N \times N$ algorithm-specific positive definite (kernel) matrix constructed from the data points [14,36], and $\boldsymbol{1}$ is an $N \times 1$ vector of ones. The first constraint in (4) ensures that the projections have zero means. The last two constraints restrict the projections to be orthonormal. The solution to Problem (4) is given by the d top eigenvectors of the centered kernel matrix $(\boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{K}(\boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T)$. When $\boldsymbol{K}$ is a stochastic matrix (e.g. LLE, LEM), the solution is simply eigenvectors $2, \ldots, d+1$ of $\boldsymbol{K}$ (without centering).

The orthogonality constraints in (4) guarantee that the projections be linearly independent. However, they do not guarantee non-redundancy. To see this, it is insightful to interpret them in statistical terms. Assume that the data points $\{\boldsymbol{x}_n\}$ correspond to independent realizations of some random vector X. Then orthogonality corresponds to zero statistical correlation, as

$$\mathbb{E}\left[f_i(X)f_j(X)\right] \approx \tfrac{1}{N}\sum_n f_i(\boldsymbol{x}_n)f_j(\boldsymbol{x}_n) = \tfrac{1}{N}\boldsymbol{f}_i^T \boldsymbol{f}_j = 0. \tag{5}$$

Therefore, in particular, the constraints in (4) guarantee that each projection be uncorrelated with any linear combination of the preceding projections, so that

---

[1] Note that LEM and DFM rather use *weighted* orthogonality constraints, but they can also be brought into the form of (4) (see supplementary material). Also, note that some methods (e.g. LEM, LLE) rather *minimize* the objective in (4). These problems can be cast as (4) with the kernel $\check{\boldsymbol{K}} = \lambda_{\max}\boldsymbol{I} - \boldsymbol{K}$ [4,17].

$f_i(X)$ *cannot be a linear function of the previous projections* $\{f_j(X)\}_{j<i}$. However, these constraints do not prevent $f_i(X)$ from being a *nonlinear* function of the previous projections, which would lead to redundancy, as in Figs. 1, 2 and 4.

To enforce non-redundancy, i.e. that each projection is not a function of the previous ones, we propose to use the following observation.

Lemma 1. *A sequence of non-trivial zero-mean projections* $\{f_i\}$ *is **non-redundant** if each of them is **unpredictable** from the preceding ones, namely*

$$\mathbb{E}\left[f_i(X)|f_{i-1}(X),\cdots,f_1(X)\right] = 0. \qquad (6)$$

Proof. Assume (6) holds and suppose to the contrary that the ith projection is non-trivial and redundant, so that $f_i(X) = h(f_{i-1}(X),\ldots,f_1(X))$ for some function h. From the orthogonality property of the conditional expectation,

$$\mathbb{E}[(f_i(X) - \mathbb{E}[f_i(X)|f_{i-1}(X),\cdots,f_1(X)])\,g(f_{i-1}(X),\cdots,f_1(X))] = 0 \qquad (7)$$

for every function g. Substituting (6), this property implies that

$$\mathbb{E}\left[f_i(X)\,g(f_{i-1}(X),\cdots,f_1(X))\right] = 0, \quad \forall g. \qquad (8)$$

Therefore, in particular, for $g \equiv h$ we get that $\mathbb{E}[f_i^2(X)] = 0$, contradicting our assumption that $f_i(X)$ is non-trivial. □

Notice that by enforcing unpredictability, we in fact restrict each projection to be uncorrelated with *any function* of the previous projections (see (8)). This constraint is stronger than the original zero correlation constraint (5), yet less restrictive than statistical independence. Specifically, two random variables Y, Z are independent if and only if $\mathbb{E}\left[g(Y)h(Z)\right] = \mathbb{E}\left[g(Y)\right]\mathbb{E}\left[h(Z)\right], \forall g, h$, whereas for Y to be unpredictable from Z it is only required that $\mathbb{E}\left[Y\,h(Z)\right] = \mathbb{E}\left[Y\right]\mathbb{E}\left[h(Z)\right], \forall h$ (corresponding to (8) in the case of zero-mean variables).

4   Algorithm

The unpredictability condition (6) is in fact an infinite set (a continuum) of constraints, as it restricts the conditional expectation of $f_i(X)$ to be zero, given every possible value that the previous projections $\{f_j(X)\}_{j<i}$ may take. However, in practice, spectral methods compute the projections only at the sample points. Therefore, to obtain a practical method, we propose to enforce these restrictions only at the sample embedding points, leading to a discrete set of N constraints

$$\mathbb{E}\left[f_i(X)|\{f_j(X) = f_j(\boldsymbol{x}_n)\}_{j<i}\right] = 0, \quad n = 1,\ldots,N. \qquad (9)$$

These N conditional expectations can be approximated using a kernel smoother matrix $\boldsymbol{P}_i \in \mathbb{R}^{N \times N}$ for regressing $\boldsymbol{f}_i$ against $\boldsymbol{f}_{i-1},\ldots,\boldsymbol{f}_1$, so that the nth entry of the vector $\boldsymbol{P}_i\boldsymbol{f}_i$ approximates the nth conditional expectation in (9),

$$[\boldsymbol{P}_i\boldsymbol{f}_i]_n \approx \mathbb{E}\left[f_i(X)|\{f_j(X) = f_j(\boldsymbol{x}_n)\}_{j<i}\right]. \qquad (10)$$

When using the Nadaraya-Watson estimator [27,39], the accuracy of this approximation is $\mathcal{O}(N^{-4/(i+3)})$. We therefore propose to replace the zero-correlation constraints $\boldsymbol{f}_i^T \boldsymbol{f}_j = 0$ in (4), by the unpredictability restrictions $\boldsymbol{P}_i \boldsymbol{f}_i = \boldsymbol{0}$. Our proposed redundancy-avoiding version of the spectral dimensionality reduction problem (4) is thus

Fig. 5. Top 100 of 15K singular values of the matrix $\boldsymbol{P}_2$ in the MNIST experiment of Fig. 3. The matrix is very close to being low-rank: 0.1% of its singular values account for over 99.9% of its Frobenius norm.

$$
\begin{aligned}
\max_{\boldsymbol{f}_i} \quad & \boldsymbol{f}_i^T \boldsymbol{K} \boldsymbol{f}_i \\
\text{s.t.} \quad & \boldsymbol{1}^T \boldsymbol{f}_i = 0 \\
& \boldsymbol{f}_i^T \boldsymbol{f}_i = 1 \\
& \boldsymbol{P}_i \boldsymbol{f}_i = \boldsymbol{0}, \qquad \forall i > 1.
\end{aligned} \tag{11}
$$

In the continuous domain, the conditional expectation operator has a non-empty null space. However, this property is usually not maintained by non-parametric sample approximations, like kernel regressors. As a result, the matrix $\boldsymbol{P}_i$ will typically be only *approximately* low-rank. Figure 5 shows a representative example, where 0.1% of the singular values account for over 99.9% of the Frobenius norm. To ensure that $\boldsymbol{P}_i$ is strictly low-rank (so that $\boldsymbol{P}_i \boldsymbol{f}_i = \boldsymbol{0}$ is not an empty set), we truncate its negligible singular values.
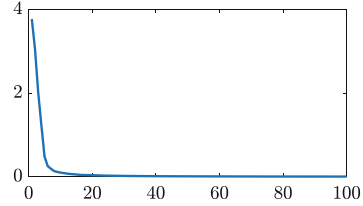
The solution to problem (11) is no longer given by the spectral decomposition of $\boldsymbol{K}$. However, it can be brought into a convenient form by using the following lemma[2] (see proof in Appendix A).

Lemma 2. *Denote the compact SVD of $\boldsymbol{P}_i$ by $\boldsymbol{U}_i \boldsymbol{D}_i \boldsymbol{V}_i^T$. Then the vectors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d$ which optimize Problem (11), also optimize*

$$
\begin{aligned}
\max_{\boldsymbol{f}_i} \quad & \boldsymbol{f}_i^T \tilde{\boldsymbol{K}}_i \boldsymbol{f}_i \\
\text{s.t.} \quad & \boldsymbol{1}^T \boldsymbol{f}_i = 0 \\
& \boldsymbol{f}_i^T \boldsymbol{f}_i = 1,
\end{aligned} \tag{12}
$$

where $\tilde{\boldsymbol{K}}_i = (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T)$ and $\boldsymbol{V}_1 = \boldsymbol{0}$.

From this lemma, it becomes clear that $\boldsymbol{f}_i$ is precisely the top eigenvector of $\tilde{\boldsymbol{K}}_i$. This implies that we can determine the non-redundant projections sequentially. In the ith step, we first modify the kernel $\boldsymbol{K}$ according to the previous projections $\boldsymbol{f}_{i-1}, \ldots, \boldsymbol{f}_1$ to obtain $\tilde{\boldsymbol{K}}_i$. Then, we compute its top eigenvector to obtain projection $\boldsymbol{f}_i$. This is summarized in Algorithm 1, where for concreteness, we chose $\boldsymbol{P}_i$ to be the Nadaraya-Watson smoother with a Gaussian-kernel.

---

[2] Note that this lemma holds true only for *maximization* problems.

Algorithm 1. Non-redundant dimensionality reduction.

Input: High-dimensional data points $\boldsymbol{x}_n \in \mathbb{R}^D$.
Output: Embeddings $\boldsymbol{f}_i = (f_i(\boldsymbol{x}_1), \cdots, f_i(\boldsymbol{x}_N))^T$.
 1: Construct the kernel matrix $\boldsymbol{K}$ as in the original algorithm (e.g. LLE, LEM, Isomap, etc.).
 2: If the original algorithm *minimizes* the objective of (4) (e.g. LLE, LEM), then set $\boldsymbol{K} \leftarrow \lambda_{\max} \boldsymbol{I} - \boldsymbol{K}$.
 3: Assign the top (non-trivial) eigen-vector of $\boldsymbol{K}$ to $\boldsymbol{f}_1$.
 4: **for** $i = 2, \ldots, d$ **do**
 5:     Construct smoothing matrix

$$[\boldsymbol{P}_i]_{j,k} \leftarrow \exp\left\{-\tfrac{1}{2h^2} \sum_{\ell=1}^{i-1} \left(f_\ell(x_j) - f_\ell(x_k)\right)^2\right\}, \quad [\boldsymbol{P}_i]_{j,k} \leftarrow \frac{[\boldsymbol{P}_i]_{j,k}}{\sum_{n=1}^N [\boldsymbol{P}_i]_{j,n}}.$$

 6:     Compute $\boldsymbol{V}_i \in \mathbb{R}^{N \times r}$, the top r right singular vectors of $\boldsymbol{P}_i$ accounting for all non-negligible singular values.
 7:     Form the modified kernel matrix

$$\tilde{\boldsymbol{K}}_i \leftarrow \left(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T\right) \boldsymbol{K} \left(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T\right).$$

 8:     Assign the top eigen-vector of $\tilde{\boldsymbol{K}}_i$ to $\boldsymbol{f}_i$.
 9: **end for**

Efficient Implementation. We use the fast method of [16] to compute the top eigenvector of $\tilde{\boldsymbol{K}}_i$ (step 8). Each iteration of [16] involves multiplication by $\tilde{\boldsymbol{K}}_i$, which can be broken into efficient multiplications by $\boldsymbol{V}_i$ and $\boldsymbol{V}_i^T$ which are $N \times r$ and $r \times N$ with $r \ll N$, and by $\boldsymbol{K}$ which is usually sparse by construction (e.g. in LEM, LLE, LTSA). Thus, we never explicitly form the matrix $\tilde{\boldsymbol{K}}_i$ (step 7).

When memory resources are restrictive, we construct a *sparse* smoothing matrix $\boldsymbol{P}_i$ (step 5) by using only the k nearest neighbors of each sample. To minimize the degradation in the representation quality we use the maximal k such that $\boldsymbol{P}_i$ fits in memory.

4.1   Relation to Independent Component Analysis (ICA)

Our method may seem similar to ICA [19,22], however, they are quite distinct. First, the ICA *objective* is independence (without preservation of geometrical structure), while in our method the objective is to preserve geometric structure subject to a statistical constraint on the embedding coordinates. Second, *non-linear* ICA is an under-determined problem, making it necessary to impose assumptions or to restrict the class of non-linear functions [20,33]. Finally, independence is a stronger constraint than unpredictability, and would thus narrow the set of possible solutions. This is while, as we saw, unpredictability is enough for avoiding redundancy.

5   Experiments

We tested our non-redundant algorithm on three high-dimensional data sets. In all our experiments, we report results with the Nadaraya-Watson smoother [27,39], as specified in Algorithm 1. We also experimented with a locally linear smoother and did not observe a significant difference. The kernel smoother bandwidth h was set adaptively: for computing $\boldsymbol{P}_i$, we took $h = \alpha(\sum_{j=1}^{i-1} \frac{1}{N}\|\boldsymbol{f}_j\|^2)^{1/2}$, where the parameter $\alpha \in [0.1, 0.6]$ was chosen using a tune set in the classification task and manually in the visualization tasks. Singular vectors corresponding to singular values smaller than 3% of the largest singular value were truncated. We used the largest number of nearest neighbors such that $\boldsymbol{P}_i$ could still be stored in memory (10K in our case). A hyper-parameter analysis is included in the supplementary material.

5.1   Artificial Head Images

The artificial head image dataset [35] is a popular test bed for manifold learning techniques. It contains 64×64 computer-rendered images of a head, with varying vertical and horizontal camera positions (denoted by θ and ϕ) and lighting directions (denoted by ψ). Linear methods (e.g. PCA, ICA) fail to detect these underlying parameters [35]. However, most (non-linear) spectral methods manage to non-redundantly extract those parameters with the first three projections, since each of the parameters (θ, ϕ, ψ) varies significantly across this data set.

   Here, to make the representation learning task more challenging, we chose a 257 subset of the original data set, corresponding to the reduced parameter range $\theta \in [-75°, 75°]$, $\phi \in [-8°, 8°]$, $\psi \in [105°, 175°]$. Figures 6(a), (c) visualize the projections extracted by LEM and LTSA in this case. As can be seen, both algorithms produce redundant representations, as their second projection is a deterministic function of the first. When incorporating our unpredictability constraints, we are able to avoid this repetition and to reveal additional information with the second projection, as evident from Figs. 6(b), (d). We quantify this by reconstructing the images from their two-dimensional embeddings using leave-one-out prediction with a non-parametric regressor (Nadaraya-Watson [27,39]). The average reconstruction peak signal to noise ratio (PSNR) is 18.0/18.2 for the original LEM/LTSA and 19.2/19.9 with our non-redundant LEM/LTSA.

   To analyze what the projections capture, we plot in Fig. 6(e)–(h) each of the embedding coordinates vs. the horizontal and vertical camera positions. From Figs. 6(e), (g) we see that in the original algorithms, Projections 1 and 2 are both correlated only with the horizontal angle θ. In our approach, on the other hand, Projection 1 captures the horizontal angle θ while Projection 2 reveals the vertical angle ϕ (see Figs. 6(f), (h)).

5.2   Image Patch Representation

To visualize the effect of non-redundancy in low-level vision tasks, we extracted all 7×7 patches with 3 pixel overlap from an image (taken from [31]), and learned
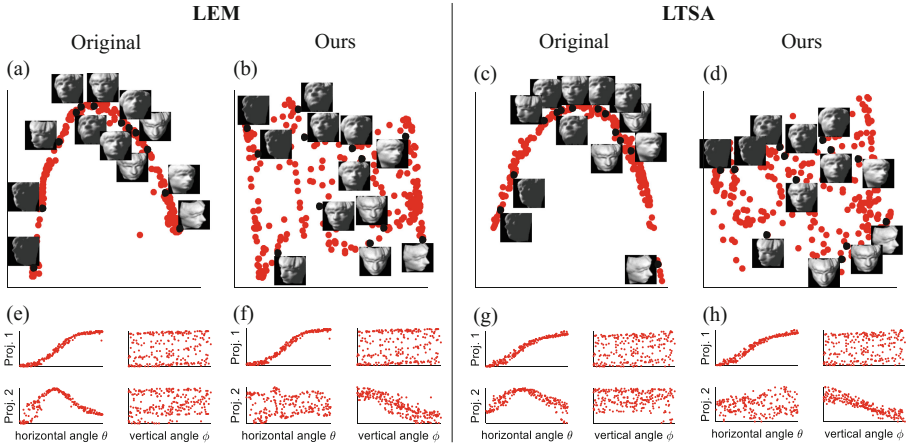
Fig. 6. Two-dimensional embeddings of computer rendered head images with varying pose and lighting directions. (a) The original LEM method. (b) Our non-redundant LEM. (c) The original LTSA method. (d) Our non-redundant LTSA. In the original algorithms, the second coordinate is a function of the first. In our method, the second coordinate clearly carries additional information w.r.t. first, and is thus non-redundant. (e)–(h) The first two projections of the head images vs. the horizontal and vertical angles (θ, ϕ) of the heads. The two projections extracted by the original algorithms are *both* correlated only with the horizontal angle θ. In our non-redundant algorithms, on the other hand, the second projection is correlated with the vertical angle ϕ.

a three dimensional representation using Isomap and using our non-redundant version of Isomap. Figure 7 visualizes the first three projections by coloring each pixel according to the embedding value of its surrounding patch. Observe that in the original algorithm, the first two projections redundantly capture brightness attributes, and the third captures mainly vertical edges with some brightness attributes still remaining (e.g. the sky, the left poolside). In contrast, in our algorithm, the second and third projections capture the vertical *and horizontal* edges (without redundantly capturing brightness multiple times), thus providing additional information. The redundancy of the 2nd Isomap projection can be seen in the scatter plot of the 2nd projection vs. the 1st. With our non-redundant algorithm, the 2nd projection is clearly not a function of the 1st, and thus captures new informative features. To quantify the amount of redundancy, we reconstruct the patches from their three-dimensional embeddings using leave-one-out prediction with the Nadaraya-Watson regressor, and then form an image by averaging overlapping patches (Fig. 8). The reconstruction PSNR is 32.9 using Isomap and 33.2 using our non-redundant Isomap.

Notice that the brightness and gradient features are linear functions of the input patches. Thus, our extracted 3D manifold is in fact linear and would be also correctly revealed by linear methods, such as PCA (not shown). Nevertheless, Isomap which is a nonlinear method, fails to extract this linear manifold due to redundancy (similarly to Fig. 2). In contrast, our non-redundant algorithm can reveal the underlying manifold regardless of its complexity.
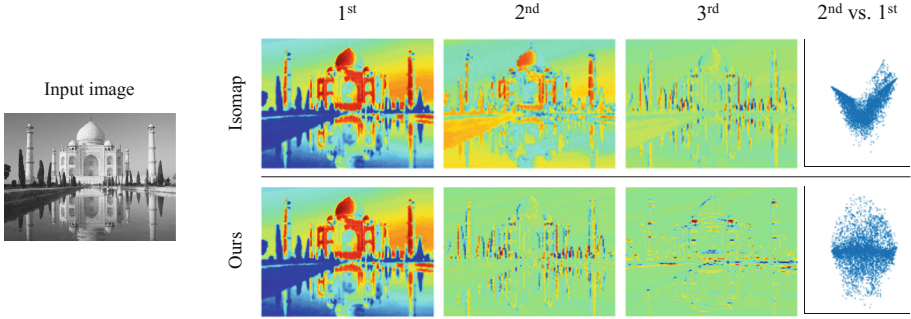
Fig. 7. Three-dimensional embedding of all 7×7 patches with a 3 pixel overlap, obtained with Isomap and with our non-redundant version of Isomap. Each pixel is colored according to the projection of its surrounding patch. In both methods, the first projection captures brightness. However, the original Isomap redundantly captures brightness-related features again with the second projection, and captures vertical edges only with the third projection. In contrast, our non-redundant version captures vertical *and horizontal* edges with the second and third projections. The scatter plot reveals that in the original Isomap, the 2*nd* projection is a function of the 1st, while in ours it is not. (Color figure online)
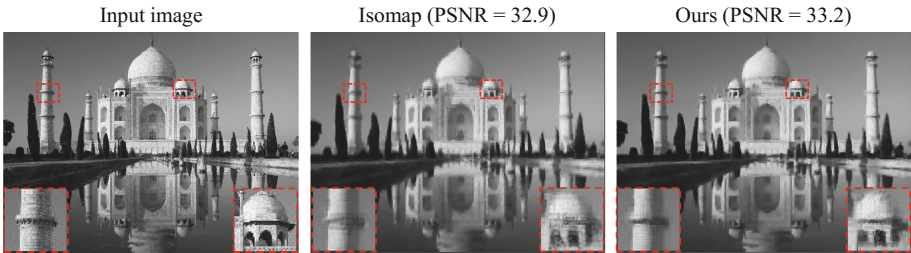
Fig. 8. The image of Fig. 7 reconstructed from the 3-dimensional patch embeddings obtained with Isomap and with our non-redundant Isomap. Note how horizontal edges are not preserved by the Isomap projections, but are preserved by our method.

5.3 MNIST Handwritten Digits

In most applications, the "correct" parametrization of the data manifold is not as obvious as in the head experiment. One example is the MNIST database [23], which contains 28×28 images of handwritten digits. In such settings, determining the quality of a low-dimensional representation can be done by measuring its impact on the performance in downstream tasks, like classification.

In the next experiment, we randomly chose a subset of 15K images from the MNIST data set, based on which we learned low-dimensional representations with LEM and with three modifications of LEM: (i) the sequential regression technique (Sect. 2), (ii) the algorithm of Dsilva et al. [10], and (iii) our non-redundant method. We then split the data into 10K/2.5K/2.5K

for training/tuning/testing and trained a third degree polynomial-kernel SVM [6] to classify the digits based on their low-dimensional representations. The SVM's soft margin parameter c and kernel parameter γ were tuned based on performance on the tune set (within the range $c \in [1, 10]$, $\gamma \in [0.1, 0.2]$). Table 1 shows the classification error for various representation sizes. As can be seen, our non-redundant representation leads to the largest and most consistent decrease in the classification error. Notice that the *linear* PCA/ICA[3] baselines are inferior in this highly non-linear scenario.

Table 1. MNIST experiment classification errors [%].

# of proj.	15K examples all labeled						# of proj.	15K examples 300 labeled	
	LEM	Dsilva et. al.	Sequential regression	Ours	PCA	ICA		LEM	Ours
3	17.6	17.6	17.3	**12.0**	47.8	59.0	5	12.6	10.3
5	8.8	8.8	14.4	**7.6**	25.7	34.1	16	8.4	6.6
7	6.9	6.9	14.2	**6.0**	14.2	19.7	24	7.2	7.2
9	6.5	6.5	14.2	**5.6**	9.9	12.9	35	7.8	8.1
11	6.0	5.4	13.8	**5.0**	7.7	6.6	50	8.8	8.8

To demonstrate the importance of *compact* representations, particularly in the semi-supervised scenario, we repeated the experiment where only 300 of the examples are labeled for the SVM training (right pan of Table 1). Notice that the error reaches a minimum at 16/24 projections with our/LEM method, and then begins to rise as the representation dimension increases. This illustrates that unnecessarily large representations result in inferior performance in downstream tasks. Our method, which is designed to construct compact representations, achieves a lower minimal error (6.6% vs. 7.2%).

Run-time. Attaining the non-redundant projections comes at the expense of increased run-time. For example, obtaining 11 projections of 15K MNIST examples takes 14 min on a 4-core Intel i5 desktop with 16 GB RAM, whereas obtaining the original LEM projections takes 13 s.

6   Conclusions

We presented a general approach for overcoming the redundancy phenomenon in spectral dimensionality reduction algorithms. As opposed to prior attempts, which fail in complex high-dimensional situations, our approach provably produces non-redundant representations. This is achieved by replacing the orthogonality constraints underlying spectral methods, by unpredictability constraints.

---

[3] In ICA, the number of independent components is equal to the dimension of the data. To obtain a low-dimensional embedding, we applied ICA on the low-dimensional embedding produced by PCA [1].

Our solution reduces to applying a sequence of spectral decompositions, where in each step, the kernel matrix is modified according to the projections computed so far. Our experiments illustrate the ability of our method to capture more informative compact representations of high-dimensional data.

A   A Proof of Lemma 2

We start by proving that any $\boldsymbol{f}_i$ solving (12) necessarily satisfies $\boldsymbol{P}_i \boldsymbol{f}_i = \boldsymbol{0}$. First, note that this constraint is equivalent to $\boldsymbol{V}_i \boldsymbol{V}_i^T \boldsymbol{f}_i = \boldsymbol{0}$, since $\boldsymbol{D}_i$, $\boldsymbol{U}_i$, and $\boldsymbol{V}_i$ have empty null spaces. Now, suppose that $\boldsymbol{f}_i$ maximizes the objective of (12) and satisfies the constraints $\|\boldsymbol{f}_i\| = 1$ and $\mathbf{1}^T \boldsymbol{f}_i = \boldsymbol{0}$, but does not satisfy $\boldsymbol{V}_i \boldsymbol{V}_i^T \boldsymbol{f}_i = \boldsymbol{0}$. Then define the alternative solution

$$\tilde{\boldsymbol{f}}_i = \frac{(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{f}_i}{\|(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{f}_i\|}, \tag{13}$$

which clearly satisfies the constraints $\|\tilde{\boldsymbol{f}}_i\| = 1$ and $\mathbf{1}^T \tilde{\boldsymbol{f}}_i = \boldsymbol{0}$, but additionally also satisfies $\boldsymbol{V}_i \boldsymbol{V}_i^T \boldsymbol{f}_i = \boldsymbol{0}$. Notice that $\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T$ is a projection matrix (as $\boldsymbol{V}_i$ is orthogonal), so that $(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T)^2 = \boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T$ and $\|(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{f}_i\|^2 \leq \|\boldsymbol{f}_i\|^2 = 1$. Therefore,

$$\tilde{\boldsymbol{f}}_i^T \tilde{\boldsymbol{K}}_i \tilde{\boldsymbol{f}}_i = \frac{\boldsymbol{f}_i^T (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T)^2 \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T)^2 \boldsymbol{f}_i}{\|(\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{f}_i\|^2} \geq \boldsymbol{f}_i^T \tilde{\boldsymbol{K}}_i \boldsymbol{f}_i, \tag{14}$$

with equality only when $\boldsymbol{V}_i \boldsymbol{V}_i^T \boldsymbol{f}_i = \boldsymbol{0}$. In other words, $\tilde{\boldsymbol{f}}_i$ achieves a higher objective value than $\boldsymbol{f}_i$, contradicting our assumption that $\boldsymbol{f}_i$ is a solution to (12). This proves that any $\boldsymbol{f}_i$ that solves problem (12) necessarily also satisfies the constraints of problem (11). Therefore, effectively, the solutions to (11) and (12) satisfy the same constraints.

Next, observe that if $\boldsymbol{f}_i$ satisfies the constraint $\boldsymbol{V}_i \boldsymbol{V}_i^T \boldsymbol{f}_i = \boldsymbol{0}$ then the objectives of (11) and (12) are equivalent, since

$$\boldsymbol{f}_i^T \tilde{\boldsymbol{K}}_i \boldsymbol{f}_i = \boldsymbol{f}_i^T (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{K} (\boldsymbol{I} - \boldsymbol{V}_i \boldsymbol{V}_i^T) \boldsymbol{f}_i = \boldsymbol{f}_i^T \boldsymbol{K} \boldsymbol{f}_i. \tag{15}$$

Therefore, $\boldsymbol{f}_i$ solves (12) if and only if it solves (11).

References

1. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. IEEE Trans. Neural Netw. **13**(6), 1450–1464 (2002)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
3. Belkin, M., Niyogi, P.: Semi-supervised learning on Riemannian manifolds. Mach. Learn. **56**(1–3), 209–239 (2004)

4. Bengio, Y., Paiement, J.F., Vincent, P., Delalleau, O., Le Roux, N., Ouimet, M.: Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In: Advances in Neural Information Processing Systems (NIPS) 16, pp. 177–184 (2004)
5. Brun, A., Park, H.-J., Knutsson, H., Westin, C.-F.: Coloring of DT-MRI fiber traces using Laplacian eigenmaps. In: Moreno-Díaz, R., Pichler, F. (eds.) EUROCAST 2003. LNCS, vol. 2809, pp. 518–529. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45210-2_47
6. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
7. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, p. I (2004)
8. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006)
9. Donoho, D.L., Grimes, C.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. **100**(10), 5591–5596 (2003)
10. Dsilva, C.J., Talmon, R., Coifman, R.R., Kevrekidis, I.G.: Parsimonious representation of nonlinear dynamical systems through manifold learning: a chemotaxis case study. Appl. Comput. Harmon. Anal. (2015, in press)
11. Elgammal, A., Lee, C.S.: Inferring 3D body pose from silhouettes using activity manifold learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II–681 (2004)
12. Geng, X., Zhan, D.C., Zhou, Z.H.: Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **35**(6), 1098–1107 (2005)
13. Gerber, S., Tasdizen, T., Whitaker, R.: Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps. In: International Conference on Machine Learning (ICML), pp. 281–288 (2007)
14. Goldberg, Y., Zakai, A., Kushnir, D., Ritov, Y.: Manifold learning: the price of normalization. J. Mach. Learn. Res. **9**, 1909–1939 (2008)
15. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Trans. Image Process. **17**(7), 1178–1188 (2008)
16. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
17. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A kernel view of the dimensionality reduction of manifolds. In: International Conference on Machine Learning (ICML), p. 47 (2004)
18. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using Laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 328–340 (2005)
19. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. **10**(3), 626–634 (1999)
20. Hyvärinen, A., Pajunen, P.: Nonlinear independent component analysis: existence and uniqueness results. Neural Netw. **12**(3), 429–439 (1999)
21. Isola, P., Zoran, D., Krishnan, D., Adelson, E.H.: Crisp boundary detection using pointwise mutual information. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 799–814. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_52

22. Jutten, C., Herault, J.: Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Sig. Process. **24**(1), 1–10 (1991)
23. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
24. Lee, C.S., Elgammal, A.: Modeling view and posture manifolds for tracking. In: International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
25. Lim, H., Camps, O.I., Sznaier, M., Morariu, V.I.: Dynamic appearance modeling for human tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 751–757 (2006)
26. Lim, I.S., de Heras Ciechomski, P., Sarni, S., Thalmann, D.: Planar arrangement of high-dimensional biomedical data sets by Isomap coordinates. In: IEEE Symposium on Computer-Based Medical Systems, pp. 50–55 (2003)
27. Nadaraya, E.A.: On estimating regression. Theory Probab. Appl. **9**(1), 141–142 (1964)
28. Pless, R.: Image spaces and video trajectories: using Isomap to explore video sequences. In: International conference on Computer Vision (ICCV), vol. 3, pp. 1433–1440 (2003)
29. Raytchev, B., Yoda, I., Sakaue, K.: Head pose estimation by nonlinear manifold learning. In: International Conference on Pattern Recognition (ICPR), vol. 4, pp. 462–466 (2004)
30. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
31. Rubinstein, M., Gutierrez, D., Sorkine, O., Shamir, A.: A comparative study of image retargeting. ACM Trans. Graph. (TOG) **29**, 160 (2010)
32. Schölkopf, B., Smola, A., Müller, K.-R.: Kernel principal component analysis. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 583–588. Springer, Heidelberg (1997). https://doi.org/10.1007/BFb0020217
33. Singer, A., Coifman, R.R.: Non-linear independent component analysis with diffusion maps. Appl. Comput. Harmon. Anal. **25**(2), 226–239 (2008)
34. Souvenir, R., Zhang, Q., Pless, R.: Image manifold interpolation using free-form deformations. In: IEEE International Conference on Image Processing (ICIP), pp. 1437–1440 (2006)
35. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
36. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality reduction: a comparative review. J. Mach. Learn. Res. **10**, 66–71 (2009)
37. Vlachos, M., Domeniconi, C., Gunopulos, D., Kollios, G., Koudas, N.: Non-linear dimensionality reduction techniques for classification and visualization. In: International Conference on Knowledge Discovery and Data Mining, pp. 645–651 (2002)
38. Wang, Q., Xu, G., Ai, H.: Learning object intrinsic structure for robust visual tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II-227 (2003)
39. Watson, G.S.: Smooth regression analysis. Sankhyā: Indian J. Stat. Ser. A **26**(4), 359–372 (1964)
40. Zhang, Z., Chow, T.W., Zhao, M.: Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization. IEEE Trans. Knowl. Data Eng. **25**(5), 1148–1161 (2013)
41. Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. J. Shanghai Univ. **8**(4), 406–424 (2004)