Semi-supervised Riemannian Dimensionality Reduction and Classification Using a Manifold-based Random Walker Graph

Faezeh Fallah, Felix Wiewel, Bin Yang
Institute of Signal Processing and System Theory, University of Stuttgart
Stuttgart, Germany
https://orcid.org/0000-0002-7115-9161

***Abstract*—Efficient classification of manifold-based data demands dimensionality reduction. However, the optimal lower dimension is mostly unknown. Also, supervised classification demands a collection of expert-annotated samples. This collection is tedious, costly, and error prone. Thus, semi-supervised classifications are motivated. In this paper, we propose a principled method to determine the optimal lower dimension and to reduce dimension of manifold-based data for their semi-supervised classification. This method relies on a manifold-based random walker graph and enhances the discriminative power of the classifier in the reduced dimensional space. The reduced dimensional data can be classified by any method or the same graph after adapting edge weights of the graph to them. Thus, the proposed dimensionality reduction is independent of the classifier type. This method is evaluated on segmenting tissues on fat-water (2-channel) magnetic resonance images.**

***Index Terms*—Riemannian dimensionality reduction, semi-supervised classification, random walker graph, symmetric positive definite matrices, region covariance descriptors.**

I. Introduction

Natural data normally lie on manifolds that do not obey Euclidean geometry. Dominant examples are region covariance descriptors (RCDs) that fuse various features to capture sophisticated regional patterns. These descriptors are symmetric positive definite (SPD) matrices that lie on a Riemannian manifold embedded in a high-dimensional Euclidean space. Data processing in this space has high complexities and can lead to the curse of dimensionality [1], [2]. Accordingly, dimension reduction is an essential step towards processing of the manifold-based data. So far, several supervised and unsupervised methods have been proposed to reduce the dimension of such data [1], [2]. However, to the best of our knowledge, none of them provides a principled way to find the optimal lower dimension. This dimension is mostly unknown and depends on the data and the classification/clustering problem.

Neighborhood graphs such as the random walker graphs have been widely used for semi-supervised classifications [3], [4]. However, their utility is not restricted to classification problems. Belkin et al. have shown that if the manifold-based samples reside on a compact Riemannian manifold and their number tends to infinity, then there is a connection between the discrete spectrum of the Laplace-Beltrami operator on the manifold and the eigenvectors of the Laplacian matrix of a neighborhood graph built from these samples [5]. This suggests the utility of the neighborhood graphs to identify the optimal lower dimension of the manifold-based data. However, eigenmaps of these Laplacian matrices cannot reduce a curved manifold to another curved manifold. They rather reduce the manifold to one of its Hilbert subspaces. Recent advances on optimizations over Stiefel and Grassmannian manifolds allow such a nonlinear dimensionality reduction [2].

TABLE I
Used abbreviations

Dice coefficient	Dice
Mean symmetric surface distance	MSSD
Hausdorff distance	HSD
Vertebral bodies	VBs
Intervertebral discs	IVDs
Pericardial adipose tissue	PeAT
Epicardial adipose tissue	EpAT
Perivascular adipose tissue	PvAT
Dimension of the reduced Euclidean space	d
Uniform/Computed	u/c

We propose a principled method to determine the optimal lower dimension and to reduce the dimension of the manifold-based data for their semi-supervised classification. This method relies on a manifold-based random walker graph as a neighborhood graph and enhances the discriminative power of the classifier in the reduced dimensional space. After the dimension reduction, the classification can be done over the same graph by adapting its edge weights to the reduced dimensional data. The proposed method is evaluated on segmenting tissues on fat-water (2-channel) magnetic resonance (MR) images. A fat-water image is a volumetric fat and its associated volumetric water image recorded by a chemical-shift-encoded MR pulse sequence. Table I shows the used abbreviations. We denote scalars, functions, and histograms with small letters, vectors with small bold letters, matrices with bold capitals, tensors with underlined bold capitals, manifolds with italic capitals, and sets with fractured capitals.

II. Materials and Methods

A. Background

A semi-supervised classifier gets trained by using a set of labeled and unlabeled samples. We assume that every sample is an $n \times n$ SPD matrix, e.g. an RCD extracted from an image patch. A labeled and an unlabeled SPD matrix (sample) are denoted by $\mathbf{X}_i^l \in M^n$ and $\mathbf{X}_j^u \in M^n$,

respectively, where $M^n \subset \mathbb{R}^{n(n+1)/2}$ is a *curved* Riemannian manifold of dimension n. The labeled matrices form the set $\mathfrak{X}^l = \{\mathbf{X}_i^l \in M^n\}_{i=1}^{N_{\text{lab}}}$. The unlabeled matrices form the set $\mathfrak{X}^u = \{\mathbf{X}_j^u \in M^n\}_{j=1}^{N_{\text{unl}}}$. Then, the training data of the classifier is $\mathfrak{X} = \mathfrak{X}^l \bigcup \mathfrak{X}^u$. A SPD matrix, regardless of its label, is denoted by $\mathbf{X}_i \in \mathfrak{X}$. Every labeled matrix $\mathbf{X}_i^l$ has a label histogram h_i of C bins (classes). This histogram is derived from the reference labels of its primitives. For example, if the matrix is extracted from an image patch, then the primitives are the voxels of the patch. The label histogram of every unlabeled matrix $\mathbf{X}_i^u$ is assumed to be a unit variance normal distribution around the mode of the labels of the labeled primitives. This distribution has a full width at half maximum (FWHM) of 2.355 [6]. If no such an assumption can be made, then the FWHM get replaced by a hyperparameter. Prior to the classification, the manifold $M^n \subset \mathbb{R}^{n(n+1)/2}$ should be reduced to another curved Riemannian manifold $M^m \subset \mathbb{R}^{m(m+1)/2}$ of an optimally lower dimension $m < n$. The optimal dimension should lead to the best classification performance on unseen test samples.

Belkin et al. [5] have shown that if the manifold M^n is compact, then its Laplace-Beltrami operator has a discrete spectrum and each of its spectral elements corresponds to an eigenfunction (basis) of a linear Hilbert subspace of the manifold. That is, any function, in particular, any classifier on the Hilbert subspace can be expressed by a linear combination of these eigenfunctions. The Hilbert subspace has a dimension less than the embedding space of the manifold but is not a curved Riemannian manifold. Thus, the underlying eigenmaps cannot support a manifold-to-manifold reduction. However, if a weighted undirected neighborhood graph is built from the samples and the number of samples tends to infinity, then

- the geodesic distance between every two vertices of the neighborhood graph converges to the geodesic distance between the corresponding samples on the manifold M^n.
- every eigenvector of the Laplacian matrix of the neighborhood graph corresponds to an eigenfunction of the aforementioned Hilbert subspace of the manifold M^n.

The geodesic distance between two vertices of the neighborhood graph is the smallest sum of the edge weights among all the connecting paths between the vertices [4].

The samples $\{\mathbf{X}_i \in M^n \subset \mathbb{R}^{n(n+1)/2}\}$ are real-valued and the Hilbert subspace has a finite dimension. Thus, the Hilbert subspace is an Euclidean space denoted by $\mathbb{R}^d$. This space is spanned by d mutually orthogonal eigenvectors of the Laplacian matrix of the neighborhood graph [5]. The space $\mathbb{R}^d$ should encompass the reduced manifold $M^m \subset \mathbb{R}^{m(m+1)/2}$. That is, $\mathbb{R}^{m(m+1)/2} \subseteq \mathbb{R}^d$ or m can be deduced from d as

$$\tfrac{m(m+1)}{2} \leq d < \tfrac{n(n+1)}{2} \implies m = \lfloor (\sqrt{1+8d}-1)/2 \rfloor \quad (1)$$

B. The Manifold-based Random Walker Graph

To find the optimal dimension d, we model the manifold M^n by a weighted undirected neighborhood graph that also incorporates labels or priors of the samples. This graph expands on a feature- and prior-based random walker graph [3], by replacing features with the SPD matrices of the samples. The graph is denoted by $\mathcal{G} = \{\mathcal{G}_f, \mathcal{G}_p\}$ with $\mathcal{G}_f$ and $\mathcal{G}_p$ being the feature-based and the prior-based subgraph, respectively. The feature-based subgraph $\mathcal{G}_f = (\mathcal{V}_f, \mathcal{E}_f)$, with vertices $\mathcal{V}_f$ and edges $\mathcal{E}_f$, encodes feature differences of the *neighboring* samples. Thus, it is a *neighborhood* graph. The neighborhood can be in spatial or temporal domain. The prior-based subgraph $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$, with vertices $\mathcal{V}_p$ and edges $\mathcal{E}_p$, encodes labels or priors. It has C vertices for C classes.

In $\mathcal{G}_f$, every vertex $v_{i_f} \in \mathcal{V}_f$ represents a SPD matrix (sample) $\mathbf{X}_i \in \mathfrak{X}$. If two matrices (samples) $\mathbf{X}_i, \mathbf{X}_j \in \mathfrak{X}$ are neighbors of each other, then their vertices $v_{i_f}, v_{j_f} \in \mathcal{V}_f$ are connected via an edge $e_{ij_f} \in \mathcal{E}_f$ with a weight $w_{ij_f} \in \mathbb{R}$.

In $\mathcal{G}_p$, every vertex $v_{c_p} \in \mathcal{V}_p$ represents a class $c \in C$ and is connected to all vertices of the feature-based subgraph $\mathcal{G}_f$. The edge connecting $v_{c_p} \in \mathcal{V}_p$ with $v_{i_f} \in \mathcal{V}_f$ is $e_{ci_p} \in \mathcal{E}_p$. According to the prior-based random walker graph [3], [4], this edge has a weight $w_{ci_p} = p_{ci} \cdot \lambda_p$, where $0 \leq \lambda_p \leq 1$ is a hyperparameter and $0 \leq p_{ci} \leq 1$ is the prior probability of the c^{th} class for the sample (matrix) $\mathbf{X}_i \in \mathfrak{X}$ at $v_{i_f} \in \mathcal{V}_f$. If this sample has the reference label $c \in C$, then $p_{ci} = 1$ and $p_{c'i} = 0,\ \forall c' \neq c$. If it has no label, then its priors are used. If it has neither label nor priors, then uniform priors $\{p_{ci} = 1/C\}_{c=1}^{C}$ are assumed for it. The priors get normalized to fulfill $\sum_{c=1}^{C} p_{ci} = 1$. This way, the partially labeled samples are incorporated to train a semi-supervised classifier.

Similar to [3], we define the edge weights of the feature-based subgraph $\mathcal{G}_f$ based on a weighted Tukey's biweight function of the feature (matrix) differences between neighboring samples. This function has a parameter σ_t to enhance the classification and the outlier removal. We consider that every $n \times n$ matrix (sample) $\mathbf{X}_i \in \mathfrak{X}$ has N_{neigh} neighbors and resides on the manifold M^n. Also, the total number of samples is $N_{\text{samp}} = |\mathfrak{X}| = |\mathfrak{X}^l \bigcup \mathfrak{X}^u|$ and the q^{th} neighbor of $\mathbf{X}_i \in \mathfrak{X}$ is $\mathbf{X}_{i_q} \in \mathfrak{X}$. Based on these:

$$\mathbf{D}_{iq} = \mathbf{X}_i - \mathbf{X}_{i_q}, \quad (2)$$

$$\underline{\mathbf{D}}_{N_{\text{samp}} \times N_{\text{neigh}} \times n \times n} = \big[\mathbf{D}_{iq}\big]_{\substack{1 \leq i \leq N_{\text{samp}} \\ 1 \leq q \leq N_{\text{neigh}}}}, \quad (3)$$

$$\underline{\mathbf{M}}'_{N_{\text{neigh}} \times n \times n} = \quad (4)$$
$$= \underset{\mathfrak{X}}{\text{median}} \big[\underline{\mathbf{D}}_{N_{\text{samp}} \times N_{\text{neigh}} \times n \times n}\big],$$

$$\underline{\mathbf{M}}_{N_{\text{samp}} \times N_{\text{neigh}} \times n \times n} = \quad (5)$$
$$= \mathbf{1}_{N_{\text{samp}}} \otimes \underline{\mathbf{M}}'_{N_{\text{neigh}} \times n \times n},$$

$$\sigma_t = \sqrt{5} \times 1.4826 \times \left\| \underset{\mathfrak{X}}{\text{median}} \left[\underline{\mathbf{D}} - \underline{\mathbf{M}}\right] \right\|_F, \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{1}_{N_{\text{samp}}} \otimes (\cdot)$ replicates $(\cdot)$ by a factor of N_{samp}. The constant $\sqrt{5} \times 1.4826$ stems from a zero-mean normal distribution assumed for the features that form the SPD matrices and considering the minimum outlier removing point of the Tukey's biweight function [3].

Then we define the weight $w_{ij_f} \in \mathbb{R}$ as

$$w_{ij_f} = \begin{cases} \big[1 - (x/\sigma_t)^2\big]^2 \cdot \exp(-r_{ij}), & |x| \leq \sigma_t, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

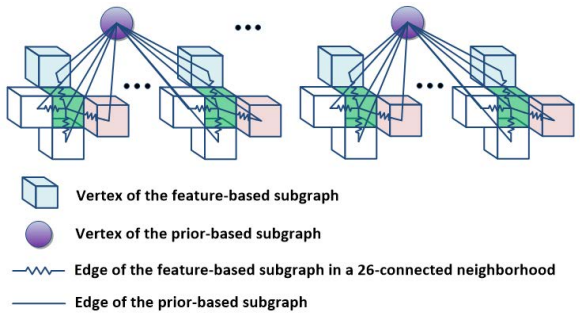

Fig. 1. Parts of the feature- and prior-based subgraphs for the image segmentation. The feature-based subgraph encodes the 26-connected (spatial) neighborhoods between the image patches represented by the cubic vertices.

where $x = \|\mathbf{X}_i - \mathbf{X}_j\|_F$ and $r_{ij} = \mathrm{FWHM}(h_i) - \mathrm{FWHM}(h_j)$ is the relative impurity of $\mathbf{X}_i$ and $\mathbf{X}_j$ defined in terms of the FWHM of their label histograms h_i and h_j. If $\mathbf{X}_i$ is unlabeled, then $\mathrm{FWHM}(h_i) = 2.355$ (see II-A).

C. Finding the Optimal Lower Dimension

The built graph has a symmetric positive semi-definite Laplacian matrix $\mathbf{L}_{N_{\mathrm{samp}} \times N_{\mathrm{samp}}}$ which is diagonalizable by its eigenvalue decomposition. In this diagonalizable matrix, the algebraic and the geometric multiplicity of every eigenvalue are equal. The semi-definiteness implies $\mathbf{L}$ to have N_{samp} (not necessarily distinct) real nonnegative eigenvalues [7]. The algebraic multiplicity of the zero eigenvalue equals the number of disconnected parts in the graph. The smallest nonzero eigenvalues define the Rayleigh quotient, the internal structure of the subgraphs, and their clustering or classification. Thus, for our analyzes, we only need to span $\mathbb{R}^d$ by d mutually orthogonal eigenvectors of these smallest nonzero eigenvalues.

Every matrix $\mathbf{X}_i \in M^n \subset \mathbb{R}^{n(n+1)/2}$ can be represented by a vector $\mathbf{x}_i \in \mathbb{R}^{n(n+1)/2}$. Thus, the dimensionality reduction should convert $\mathbf{x}_i \in \mathbb{R}^{n(n+1)/2}$ to $\mathbf{y}_i \in \mathbb{R}^d$ with a minimum reconstruction error. In this regard, to find the optimal d, we compute the eigenvalues of $\mathbf{L}$ and sort them in an ascending order. Then, we start from the smallest nonzero one. In the k^{th} iteration, we use $d^{(k)}$ eigenvectors of the selected eigenvalues to reconstruct $\{\mathbf{x}_i \in \mathbb{R}^{n(n+1)/2}\}$ with $\{\mathbf{y}_i^{(k)} \in \mathbb{R}^{d^{(k)}}\}$. Then we compute the error of these reconstructions and if this error is above a threshold ϵ we take the next eigenvalue from the sorted list and add it to the selected ones. If the error is below the threshold, the current number of eigenvectors defines d. The reconstruction error in the k^{th} iteration is

$$e^{(k)} = \mathrm{tr}(\mathbf{Y}^{(k)^T} \mathbf{L} \mathbf{Y}^{(k)}). \tag{8}$$

The i^{th} row of $\mathbf{Y}^{(k)}_{N_{\mathrm{samp}} \times d^{(k)}}$ is $\mathbf{y}_i^{(k)^T}$; $\mathrm{tr}(\cdot)$ is the trace operator. The above error with a constraint on scale-invariance forms the objective function for computing the Laplacian eigenmaps [5]. Having d determined, the dimension m of the reduced manifold can be computed using (1).

D. Semi-supervised Riemannian Dimensionality Reduction

M^n can be reduced to M^m by a map $f_{\mathbf{W}} : M^n \to M^m$ which operates with the matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ and converts every SPD matrix $\mathbf{X}_i \in M^n$ to another SPD matrix $\mathbf{W}^T \mathbf{X}_i \mathbf{W} \in M^m$. This map can address different goals. In our case, we aim to enhance the discriminative power of the semi-supervised classifier in the reduced dimensional space. For this, we combine a supervised [1] and an unsupervised [2] Riemannian dimensionality reduction and use both the labeled $\mathfrak{X}^l = \{\mathbf{X}_i^l \in M^n\}_{i=1}^{N_{\mathrm{lab}}}$ and unlabeled $\mathfrak{X}^u = \{\mathbf{X}_j^u \in M^n\}_{j=1}^{N_{\mathrm{unl}}}$ matrices to compute $\mathbf{W} \in \mathbb{R}^{n \times m}$ as

$$\begin{aligned} \mathbf{W} = \arg\min_{\hat{\mathbf{W}}} \Bigg(& \sum_{i=1}^{N_{\mathrm{lab}}} \sum_{j=1}^{N_{\mathrm{lab}}} a_l(\mathbf{X}_i^l, \mathbf{X}_j^l) \cdot \delta_J(\hat{\mathbf{W}}^T \mathbf{X}_i^l \hat{\mathbf{W}}, \hat{\mathbf{W}}^T \mathbf{X}_j^l \hat{\mathbf{W}}) \\ & + \sum_{i=1}^{N_{\mathrm{lab}}} \sum_{j=1}^{N_{\mathrm{unl}}} a_u(\mathbf{X}_i^l, \mathbf{X}_j^u) \cdot \delta_J(\hat{\mathbf{W}}^T \mathbf{X}_i^l \hat{\mathbf{W}}, \hat{\mathbf{W}}^T \mathbf{X}_j^u \hat{\mathbf{W}}) \\ & - \sum_{i=1}^{N_{\mathrm{unl}}} \delta_J(\hat{\mathbf{W}}^T \mathbf{X}_i^u \hat{\mathbf{W}}, \hat{\mathbf{W}}^T \mathbf{M} \hat{\mathbf{W}}) \Bigg) \\ & \text{subject to} \quad \hat{\mathbf{W}}^T \hat{\mathbf{W}} = \mathbf{I}_m. \end{aligned} \tag{9}$$

$\mathbf{I}_m$ is the $m \times m$ identity matrix and δ_J is the Jeffrey divergence as a Riemannian distance metric [2]. $\mathbf{M}$ is the Fréchet mean of $\mathfrak{X} = \mathfrak{X}^l \bigcup \mathfrak{X}^u$. Due to the use of the Jeffrey divergence, $\mathbf{M}$ has a closed-form analytical solution [2].

The first two terms of the above function pull together samples of low distance or similar labels and push apart samples of large distance or different labels. Thus, they enhance the discriminative power of the classifier in the reduced dimensional space. The third term maximizes the variance (distance between the unlabeled matrices and the overall Fréchet mean). This minimization is done by a conjugate gradient optimizer [8] over the Grassmannian manifold $G(m, n) \subset \mathbb{R}^{n \times m}$. The required Jacobian matrices are given in [2]. This manifold guarantees the unitary constraint $\hat{\mathbf{W}}^T \hat{\mathbf{W}} = \mathbf{I}_m$ which ensures that every $\hat{\mathbf{W}}^T \mathbf{X}_i \hat{\mathbf{W}}$ is SPD [2]. The affinity functions $a_l(\cdot)$ and $a_u(\cdot)$ encode neighborhoods and labels as

$$a_l(\mathbf{X}_i^l, \mathbf{X}_j^l) = \begin{cases} +2, & \text{if } c_i = c_j \text{ and } \mathbf{X}_i^l \leftrightarrow \mathbf{X}_j^l \\ +1, & \text{if } c_i = c_j \text{ and } \mathbf{X}_i^l \nleftrightarrow \mathbf{X}_j^l \\ -1, & \text{if } c_i \neq c_j \text{ and } \mathbf{X}_i^l \nleftrightarrow \mathbf{X}_j^l \\ -2, & \text{if } c_i \neq c_j \text{ and } \mathbf{X}_i^l \leftrightarrow \mathbf{X}_j^l \end{cases}, \tag{10}$$

TABLE II
PARAMETERS OF THE PROPOSED CLASSIFICATION FRAMEWORK WITH THE LOWER DIMENSION AT OR AROUND THE VALUE FROM II-C

Fixed/Computed Parameters					Optimized Hyperparameters							
C	N_{neigh}	d	n	m	$\sigma_t^{b,c}$	$\sigma_t^{a,c}$	$\sigma_t^{b,u}$	$\sigma_t^{a,u}$	λ_p^c	λ_p^u	ϵ^c	ϵ^u
6	26	70	96	11	7.2	6.8	6.4	5.1	2	0	11	14
6	26	80	96	12	7.1	6.5	6.1	4.9	2	0	11	12
6	26	90	96	13	6.9	6.5	5.9	4.8	2	0	9	10
6	26	110	96	14	6.5	6.3	5.9	4.7	1	0	7	9
6	26	120	96	15	6.2	5.8	5.4	4.5	1	0	6	8
6	26	134*	96	16*	5.9*	5.6*	5.1*	4.2*	1*	0*	5*	7*
6	26	150	96	17	6.1	5.7	5.3	4.4	1	0	5	8
6	26	170	96	18	6.1	5.8	5.4	4.5	1	0	6	8
6	26	190	96	19	6.3	5.9	5.7	4.8	1	0	7	8
6	26	210	96	20	6.6	6.2	5.9	5.1	1	0	7	9
6	26	230	96	21	6.7	6.2	5.9	5.2	2	0	8	10

C: Number of classes. N_{neigh}: Number of spatial neighbors.
d: Dimension of the Euclidean subspace. *: Values resulted from II-C.
n, m: Dimension of the original and the reduced manifold.
$\sigma_t^{b,c}$, $\sigma_t^{a,c}$, $\sigma_t^{b,u}$, $\sigma_t^{a,u}$: The outlier removing parameter before (b) and after (a) the dimension reduction with uniform (u) or computed (c) priors.
λ_p^c, λ_p^u: Contribution factor of the uniform (u) or computed (c) priors.
ϵ^c, ϵ^u: Maximum reconstruction error with uniform (u) or computed (c) priors.

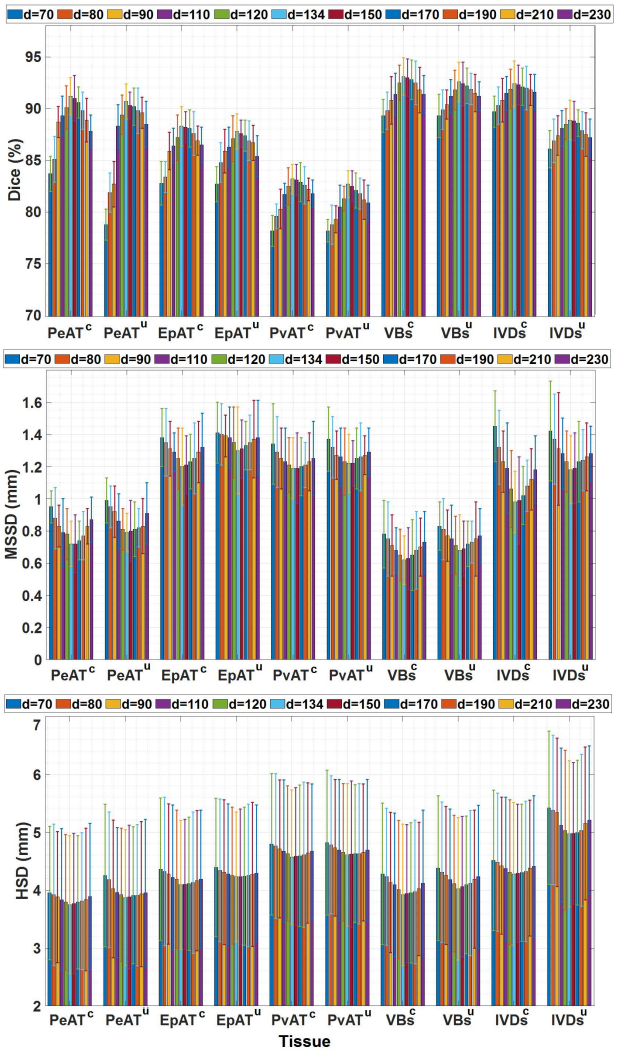

Fig. 2. Mean±SD of the objective metrics for each set of parameters associated to a d value with uniform (u) or computed (c) priors.

$$a_u(\mathbf{X}_i^l, \mathbf{X}_j^u) = \begin{cases} +1, & \text{if } \mathbf{X}_i^l \leftrightarrow \mathbf{X}_j^u \\ -1, & \text{if } \mathbf{X}_i^l \nleftrightarrow \mathbf{X}_j^u \end{cases}, \qquad (11)$$

where c_i is the mode of the label histogram of $\mathbf{X}_i^l$ and $\leftrightarrow$ ($\nleftrightarrow$) denotes matrices are (are not) neighbors of each other.

E. Classification

After reducing M^n to M^m, any classifier can classify $\{\mathbf{W}^T\mathbf{X}_i\mathbf{W} \in M^m\}$. This is because the proposed dimensionality reduction is independent from the classifier type. However, to avoid the burden of building a new classifier, the proposed random walker graph can also be used for the classifications. Edge weights of this graph reflect distances in feature space. However, the graph topology reflects neighborhoods in another (e.g. spatial or temporal) domain. Thus, after the dimension reduction, the same graph topology with updated edge weights can classify the data. This update happens by replacing $\{\mathbf{X}_i \in M^n\}$ with $\{\mathbf{W}^T\mathbf{X}_i\mathbf{W} \in M^m\}$ in (2), (3), (4), (5), (6), and (7). Then the classification probabilities of the unlabeled matrices $\{\mathbf{W}^T\mathbf{X}_i^u\mathbf{W} \in M^m\}_{i=1}^{N_{\text{unl}}}$ are solutions of the linear system of equations of the graph [9].

TABLE III
MEAN±SD OF THE OBJECTIVE METRICS FOR THE TISSUES AUTOMATICALLY SEGMENTED BY DIFFERENT METHODS

	PeAT	EpAT	PvAT	VBs	IVDs
Proposed Method with Computed Priors*					
Dice (%)	91.2 ± 1.8	88.3 ± 1.9	83.2 ± 1.4	93.1 ± 1.8	92.4 ± 2.2
MSSD (mm)	0.72 ± 0.14	1.20 ± 0.24	1.19 ± 0.19	0.62 ± 0.15	0.98 ± 0.19
HSD (mm)	3.75 ± 1.19	4.09 ± 1.11	4.57 ± 1.16	3.92 ± 1.22	4.28 ± 1.23
Proposed Method with Uniform Priors*					
Dice (%)	90.7 ± 1.7	87.8 ± 1.7	82.7 ± 1.3	92.6 ± 1.9	88.9 ± 1.9
MSSD (mm)	0.79 ± 0.12	1.30 ± 0.27	1.22 ± 0.18	0.68 ± 0.22	1.18 ± 0.21
HSD (mm)	3.87 ± 1.17	4.23 ± 1.12	4.61 ± 1.23	4.02 ± 1.23	4.97 ± 1.26
Dictionary Learning Method [1]					
Dice (%)	89.7 ± 1.1	87.9 ± 2.1	82.5 ± 1.9	91.3 ± 1.5	87.4 ± 1.2
MSSD (mm)	0.81 ± 0.11	1.24 ± 0.22	1.28 ± 0.42	0.71 ± 0.18	1.23 ± 0.19
HSD (mm)	3.86 ± 0.92	4.12 ± 1.08	4.88 ± 1.12	4.13 ± 1.27	5.05 ± 1.21
Deep Learning Method [10]					
Dice (%)	89.2 ± 1.4	86.7 ± 1.3	82.4 ± 1.7	93.1 ± 1.5	92.2 ± 1.8
MSSD (mm)	0.84 ± 0.17	1.25 ± 0.18	1.21 ± 0.15	0.60 ± 0.16	1.12 ± 0.31
HSD (mm)	3.89 ± 1.25	4.24 ± 1.31	4.73 ± 1.22	3.89 ± 1.15	4.34 ± 1.14
Graph-based Method [9]					
Dice (%)	88.2 ± 1.9	87.4 ± 2.6	81.3 ± 2.2	90.5 ± 1.9	86.2 ± 1.7
MSSD (mm)	1.1 ± 0.19	1.3 ± 0.25	1.61 ± 0.68	0.76 ± 0.21	1.25 ± 0.14
HSD (mm)	4.12 ± 1.28	4.31 ± 1.34	5.67 ± 1.52	4.21 ± 1.18	4.98 ± 1.17
Random Forest Method [11]					
Dice (%)	83.5 ± 1.4	82.6 ± 1.7	79.8 ± 1.5	92.5 ± 1.9	91.4 ± 2.3
MSSD (mm)	1.5 ± 0.22	1.8 ± 0.34	1.82 ± 0.53	0.65 ± 0.18	1.18 ± 0.42
HSD (mm)	4.86 ± 1.35	4.72 ± 1.38	5.72 ± 1.42	3.98 ± 1.12	4.61 ± 0.92

*: Results from $d = 134$ and the associated hyperparameter values (see Table II).

TABLE IV
PROCESSING TIMES OF THE PRESENT AND THE COMPARED METHODS

	Dictionary Learning [1]	Deep Learning [10]	Graph-based Method [9]	Random Forest Method [11]
Training	4.7 hrs^c	138 min^g	3.6 hrs^c	238 min^c
Segmentation	19 ± 6.2 min^c	7.4 ± 1.7 min^c	25 ± 7.5 min^c	9.2 ± 1.9 min^c
Present Method				
Prior Computation^c	**Graph Building**^{c}	**Dimension Reduction**^{c}	**Graph Updating**^{c}	**Segmentation**^{c}
27.4 min	3.9 min	4.6 min	1.7 min	28.4 ± 5.2 s

Training: Total time of building a model using 34 fat-water images.
Segmentation: Mean±SD of the time of segmenting a test (unlabeled) fat-water image.
^{g}On four NVIDIA TITAN X® GPUs each occupied by 20% of its capacity.
^{c}On a quad-core CPU of 3.10 GHz frequency and 32 GB RAM.
Prior computation is based on a multiatlas registration [4].

III. EVALUATION

The proposed dimensionality reduction method aims at enhancing the discriminative power of a classifier applied to the manifold-based data. Thus, we evaluate it on a multilabel classification task which is an automatic volumetric segmentation of multichannel (fat-water) MR images. In this application, distances (neighborhoods) in both feature and spatial domain matter. The previous dimensionality reduction methods provide no principled way to find the optimal lower dimension d [1], [2]. Thus, **to testify the optimality** of the d value from II-C, we compare the classification performances resulted from this value with the ones from values around it.

A. Image Data Sets and the Multilabel Segmentation Problem

To evaluate the proposed method, 58 volumetric fat-water images are recorded and their voxel-wise reference labels are obtained manually. From these, 44 fat-water images and their labels form the labeled set. The remaining images without their labels form the unlabeled set. In the labeled set, 34 images and their labels are used to reduce the dimension. Rest of them are used to optimize the hyperparameters. Then, dimension of the unlabeled set is reduced and edge weights of the proposed graph are adapted to the reduced dimensional data to classify them with regard to VBs, IVDs, PeATs, EpATs, PvATs, and background. Reference labels of this set allowed subjective and objective evaluations. The priors are either uniform or computed by a multiatlas registration [4].

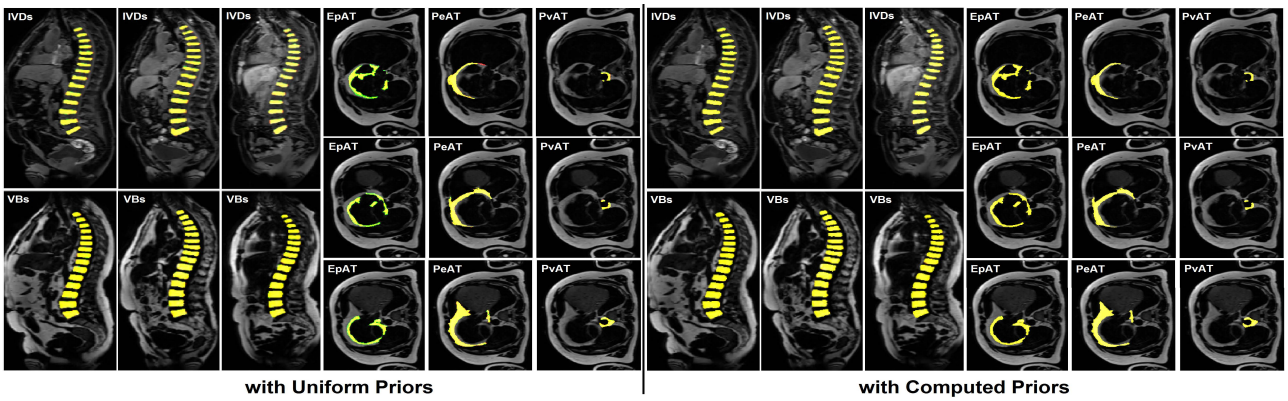

Fig. 3. Masks of the tissues automatically segmented by the proposed classification framework (red), masks of the reference labels (green), and their overlaps (yellow) on some slices of the test (unlabeled) images.

B. Dimension Reduction and Hyperparameter Optimization

On every image, a volume covering the addressed tissues is cut. In this volume, cubic patches of 27 voxels are extracted. In each patch, 90 intra-channel and 6 inter-channel features, the histogram of the reference labels, and its mode (dominant reference label) are computed. Across all the patches, each feature is normalized to zero mean and unit variance. By using the integral images of these features, the RCD of the patch is computed [1]. These RCDs are SPD matrices that lie on $M^{96} \subset \mathbb{R}^{96(96+1)/2}$. The RCDs of the labeled and unlabeled patches form $\mathfrak{X}^l = \{\mathbf{X}_i^l \in M^{96}\}_{i=1}^{34}$ and $\mathfrak{X}^u = \{\mathbf{X}_j^u \in M^{96}\}_{j=1}^{24}$, respectively. Using these sets the proposed random walker graph is built. This graph is disconnected since patches of each image are disconnected from patches of other images. Fig. 1 shows a part of this graph. On this graph, the lower dimension d is computed and the hyperparameters are cross validated by a Bayesian optimizer over randomly selected discretized parameters [12]. Table II lists these parameters. Each set (row) corresponds to a d value.

C. Automatic Volumetric Segmentations

After the classifications, for every patch, the class of highest probability defines the estimated label. These labels are compared against the dominant reference labels by using the objective metrics of Dice, MSSD, and HSD [11]. Fig. 2 shows mean$\pm$standard deviation (SD) of these metrics for each set of parameters associated to a d value. As shown, $d = 134$, computed by II-C, has led to the highest Dice and the lowest MSSD and HSD for all the classes (tissues). In Table III, this highest performance is compared against previous segmentation methods that have addressed the same tissues on the same image data set. Processing times of the compared methods are summarized in Table IV. Fig. 3 shows the segmentation results of the proposed method with $d = 134$ on some slices of the test images.

IV. Discussion and Conclusion

We presented a principled method to find the optimal lower dimension and to reduce the dimension of manifold-based data by considering their Riemannian geometry, i.e. without approximating the manifold by several affine subspaces. These allow to classify large, high-dimensional, and heterogeneous data. Even though the proposed method benefits from labels (supervision), it can be converted to an unsupervised method by replacing the labels with priors and removing the supervised term from (9). This method shares the notion of noise removal in dimension reduction with unsupervised autoencoders. However, it incorporates Riemannian geometry and explicitly aims at enhancing the discriminative power of the classifier. Future works are to incrementally incorporate every new sample instead of processing all the samples together.

References

[1] F. Fallah, K. Armanious, B. Yang, and F. Bamberg, "Spatial and Hierarchical Riemannian Dimensionality Reduction and Dictionary Learning for Segmenting Multichannel Images," in *Proc Eur Signal Process Conf*, 2019, pp. 1–5.

[2] M. Harandi, M. Salzmann, and R. Hartley, "Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 1, pp. 48–62, 2018.

[3] F. Fallah, B. Yang, S. S. Walter, and F. Bamberg, "Hierarchical feature-learning graph-based segmentation of fat-water MR images," in *Proc IEEE Conf Signal Process Algorithms Archit Arrange Appl*, 2018, pp. 37–42.

[4] F. Fallah, B. Yang, and F. Bamberg, "Automatic atlas-guided constrained random walker algorithm for 3D segmentation of muscles on water magnetic resonance images," in *Proc Eur Signal Process Conf*, 2017, pp. 251–255.

[5] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach Learn*, vol. 56, pp. 209–239, 2004.

[6] E. W. Weisstein. Full width at half maximum from mathworld. Https://mathworld.wolfram.com/GaussianFunction.html.

[7] D. A. Spielman, "Spectral graph theory," in *Combinatorial Scientific Computing*. CRC Press, 2012, ch. 18, pp. 495–524.

[8] Manopt. (2017) https://www.manopt.org/tutorial.html.

[9] F. Fallah, K. Armanious, B. Yang, and F. Bamberg, "Volumetric Surface-guided Graph-based Segmentation of Cardiac Adipose Tissues on Fat-Water MR Images," in *Proc Eur Signal Process Conf*, 2019, pp. 1–5.

[10] F. Fallah, D. M. Tsanev, B. Yang, S. Walter, and F. Bamberg, "A novel objective function based on a generalized Kelly criterion for deep learning," in *Proc IEEE Conf Signal Process Algorithms Archit Arrange Appl*, Sept 2017, pp. 84–89.

[11] F. Fallah, S. S. Walter, F. Bamberg, and B. Yang, "Simultaneous volumetric segmentation of vertebral bodies and intervertebral discs on fat-water MR images," *IEEE J Biomed Health Inform*, vol. 23, no. 4, pp. 1692–1701, 2019.

[12] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a Python library for model selection and hyperparameter optimization," *Comput Sci Discov*, vol. 8, no. 1, p. 014008, 2015.