Linear Sequence Discriminant Analysis: A Model-Based Dimensionality Reduction Method for Vector Sequences

Bing Su, Xiaoqing Ding
State Key Laboratory of Intelligent Technology and Systems
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
{subin, dxq}@ocrserv.ee.tsinghua.edu.cn

Abstract

Dimensionality reduction for vectors in sequences is challenging since labels are attached to sequences as a whole. This paper presents a model-based dimensionality reduction method for vector sequences, namely linear sequence discriminant analysis (LSDA), which attempts to find a subspace in which sequences of the same class are projected together while those of different classes are projected as far as possible. For each sequence class, an HMM is built from states of which statistics are extracted. Means of these states are linked in order to form a mean sequence, and the variance of the sequence class is defined as the sum of all variances of component states. LSDA then learns a transformation by maximizing the separability between sequence classes and at the same time minimizing the within-sequence class scatter. DTW distance between mean sequences is used to measure the separability between sequence classes. We show that the optimization problem can be approximately transformed into an eigen decomposition problem. LDA can be seen as a special case of LSDA by considering non-sequential vectors as sequences of length one. The effectiveness of the proposed LSDA is demonstrated on two individual sequence datasets from UCI machine learning repository as well as two concatenate sequence datasets: APTI Arabic printed text database and IFN/ENIT Arabic handwriting database.

1. Introduction

The targets of interest are represented with vector sequences in many computer vision and pattern recognition applications, including speech signal processing [10], online and offline handwriting recognition [3, 16], video analysis and synthesis of human motion [30]. Many of these applications focus on supervised learning, and essentially boil down to a classification problem. Depending on how sequences are associated with class labels, vector sequences can be categorized into two main types: *individual vector sequences* and *concatenate vector sequences*. Each individual vector sequence corresponds to only one pattern and can be treated as individual objects, the goal of classification is to predict a single class label for it. Each concatenate vector sequence is a concatenation of several individual vector sequences and no natural notions of segments are available. The classification task aims to detect all composed patterns and assign a label to each of them in order.

For both types of sequences, reducing the dimensionality of the vectors in sequences is necessary to discard irrelevant information and obtain more robust estimation of parameters from the computational perspective. The goal of *dimensionality reduction (DR) for vector sequences* in this paper is to map the high-dimensional vectors in sequences to a space of fewer dimensions such that discriminative information is preserved, resulting in sequences of lower dimensional vectors. The goal is different from that of discrete Fourier transform (DFT) [7] and discrete Wavelet transform (DWT) [5] based methods which operate on univariate time series and aim at reducing the length of the sequences.

Although unsupervised DR techniques such as PCA can be performed by treating all vectors equally without considering if they come from the same sequence or not, supervised methods can often achieve better performances. Various discriminant analysis techniques have been proposed [13, 11, 31], but they can not be directly applied to vector sequences for two reasons. First, it is hard to define the statistics such as mean and variance of classes since the samples are sequences, and second, the supervised information is difficult to utilize. Vectors in a sequence can neither be considered as individual samples with the same class label because they are not independent and may vary greatly, nor can they be concatenated to form a long vector since the lengths of different sequences can be different.

In this paper, we propose a model-based DR method for both types of vector sequences, namely *linear sequence discriminant analysis (LSDA)*. Statistics of sequences are obtained through model-based approach. An HMM is built for

each sequence class, mean and variance are extracted from each state of the HMM. These means are linked in order to form a sequence, which can be considered as the mean sequence of the class. Since the whole vector sequences are attached with class labels, it is hoped that sequences can be maximally separated into different classes after transformation of vectors. To achieve this objective, LSDA learns a linear transformation by maximizing the sum of pairwise dynamic time warping (DTW) distances at the same time minimizing the total between-state scatter.

The rest of this paper is organized as follows: Section 2 reviews related work on vector sequence analysis; The proposed LSDA and extended discussions are presented in section 3; Experiments and results are reported on individual sequence datasets and concatenate sequence databases in section 4 and section 5, respectively; Section 6 draws the conclusions.

2. Related work

Model-based approaches. Some statistics of vector sequences can be obtained by a category of approaches which employ dynamic system models. In these approaches [8, 26, 19], a vector sequence is considered as a series of observations generated by an underlying dynamic system, and the parameters or properties of the model can be used as some measures of statistics of the sequence class.

Hidden Markov Model (HMM) is one of the most popular dynamic system models. Based on HMM, LDA has been performed in field of speech and handwriting recognition [10, 4], which consists of two steps: 1) An HMM for each class is trained to create pseudo state labels of vectors in training sequences, 2) LDA transform is then estimated by treating all states of all HMMs as individual classes. Although statistics of sequences classes are well explored, the label information is not suitable utilized, for states within the same HMM are not independent and a true label is associated with the whole state sequence instead of a state.

DR for vector sequences. In [29, 30], DTW was combined with canonical correlation analysis to align multidimensional feature sequences. Although these methods also perform DR for vectors in sequences, they can only be applied to multi-modal sequences for alignment and can not be extended to multi-sequence classes for classification.

Kernels are often exploited in nonlinear DR techniques [27, 9]. In [21], a sequence kernel DR approach combining spatial, temporal and periodic information is proposed for time series data, where labels are associated with the vectors in long time series. The task there is to predict a class label for each frame, which is different from that of this paper to detect what patterns occur and the order of appearance in concatenate vector sequences.

Distance between sequences. For univariate time series with equal length, it is easy to define such measurements. For example, L_p norm is a natural measure of distance, and auto-correlation can also be used as measure of similarity [25]. For vector sequences with unequal length, dynamic time warping [20] is the most widely used distance measure. Given two sequences $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_x}] \in \mathbb{R}^{d\times N_x}$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_{N_y}] \in \mathbb{R}^{d\times N_y}$ with number of vectors N_x and N_y, respectively, DTW tries to find an optimal alignment from all possible sets of correspondences between vectors, along which the sum of pairwise vector-to-vector distances is minimized:

$$\min_{\pi^x,\pi^y} \sum_{t=1}^{T} \left\| \mathbf{x}_{\pi_t^x} - \mathbf{y}_{\pi_t^y} \right\|^2 \tag{1}$$

Where T is the number of steps needed to align the two sequences. $\pi^x = [\pi_1^x, \pi_2^x, \cdots, \pi_T^x]^T \in \{1 : N_x\}^{T\times 1}$ and $\pi^y = [\pi_1^y, \pi_2^y, \cdots, \pi_T^y]^T \in \{1 : N_y\}^{T\times 1}$ denote the aligned indexes between vectors in sequence $\mathbf{X}$ and $\mathbf{Y}$, respectively. Boundary conditions, continuity and monotonicity constraints are attached to π^x and π^y. Optimizing (1) can be efficiently solved by an dynamic programming (DP) algorithm.

3. Linear sequence discriminant analysis

This section presents the proposed LSDA, which considers the general C-class classification problem. The diagram of the overall process is shown in Fig. 1.

3.1. HMM-based Statistics of sequence classes

For individual sequences, each sequence class has a set of sequence samples for training. Assume that $\{\mathbf{U}_n^{\tilde{i}} | \mathbf{U}_n^{\tilde{i}} = [\mathbf{u}_{n(1)}^{\tilde{i}}, \mathbf{u}_{n(2)}^{\tilde{i}}, \cdots \mathbf{u}_{n(P_n^{\tilde{i}})}^{\tilde{i}}] \in \mathbb{R}^{d\times P_n^{\tilde{i}}}, n = 1, \cdots, N_{\tilde{i}}\}$ belong to class $\tilde{i}$, $\tilde{i} = \tilde{1}, \tilde{2}, \cdots, \tilde{C}$. $\mathbf{U}_n^{\tilde{i}}$ denote the n-th sample of class $\tilde{i}$, $\mathbf{u}_{n(p)}^{\tilde{i}}$ is the its p-th component vector. The number in subscript brackets denotes the index of the vector. $P_n^{\tilde{i}}$ is the number of vectors in the n-th sequence sample of class $\tilde{i}$. $N_{\tilde{i}}$is the number of samples of class $\tilde{i}$. Note that a tag will always be added above the class label for emphasis hereinafter.

We build a left-to-right HMM with self-loops for each sequence class. This topology can represent the evolution of sequences and is consistent with the constraints of DTW, which only allows translate to the vector itself or to the next vector for both sequences in one step. We use $\mathbf{S}^{\tilde{i}} = [\mathbf{s}_{[1]}^{\tilde{i}}, \mathbf{s}_{[2]}^{\tilde{i}}, \cdots \mathbf{s}_{[L_{\tilde{i}}]}^{\tilde{i}}]$ to denote the HMM associated with class $\tilde{i}$, where $L_{\tilde{i}}$is the number of states in $\mathbf{S}^{\tilde{i}}$, and $\mathbf{s}_{[l]}^{\tilde{i}}$ is the l-th state of $\mathbf{S}^{\tilde{i}}$. It should be noted that different models may have different number of states. Hereinafter the index of a state in the sequence model will always appears in square brackets.

As shown in Fig. 1, the sequence sample $\mathbf{U}_n^{\tilde{i}}$ of class $\tilde{i}$ is generated by the corresponding HMM $\mathbf{S}^{\tilde{i}}$, with each

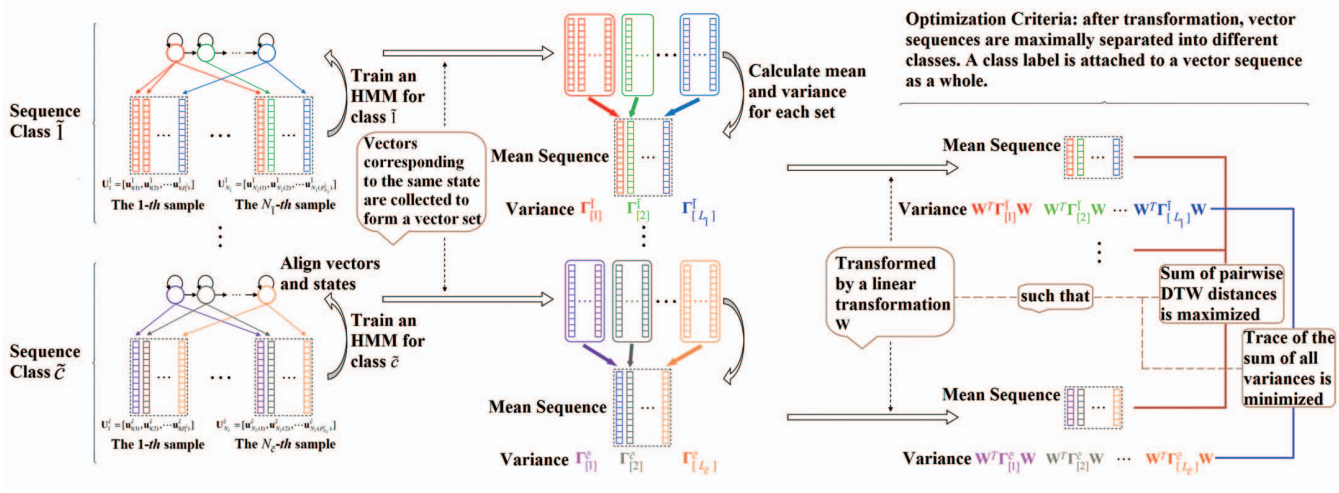


Figure 1. The diagram of the overall process of LSDA.

vector $\mathbf{u}^{\tilde{i}}_{n(p)}$ is emitted by a corresponding state $\mathbf{s}^{\tilde{i}}_{[l]}$, $l \in \{1, \cdots, L_{\tilde{i}}\}$. We put vectors in all sequence samples of class $\tilde{i}$ aligned to state $\mathbf{s}^{\tilde{i}}_{[l]}$ of $\mathbf{S}^{\tilde{i}}$ into a collection set $\mathbf{V}^{\tilde{i}}_{[l]} = \{\mathbf{u}^{\tilde{i}}_{n(p)} | \mathbf{U}^{\tilde{i}}_n \in \tilde{i}, \mathbf{u}^{\tilde{i}}_{n(p)} \in \mathbf{s}^{\tilde{i}}_{[l]}\}$, thus each state is represented with a set of vectors. Let $N^{\tilde{i}}_{[l]}$ denotes the number of vectors in this set. Statistics can then be extracted from these vector sets. We calculate mean and variance for each set. Let $\mathbf{m}^{\tilde{i}}_{[l]}$ and $\mathbf{\Gamma}^{\tilde{i}}_{[l]}$ denote the mean and variance of state $\mathbf{s}^{\tilde{i}}_{[l]}$ in $\mathbf{S}^{\tilde{i}}$, respectively.

The mean vectors $\{\mathbf{m}^{\tilde{i}}_{[l]}, l = 1, \cdots, L_{\tilde{i}}\}$ corresponding to states $\{\mathbf{s}^{\tilde{i}}_{[l]}, l = 1, \cdots, L_{\tilde{i}}\}$ of the same class model $\mathbf{S}^{\tilde{i}}$ are linked in order to form a sequence $\mathbf{m}^{\tilde{i}} = [\mathbf{m}^{\tilde{i}}_{[1]}, \mathbf{m}^{\tilde{i}}_{[2]}, \cdots, \mathbf{m}^{\tilde{i}}_{[L_{\tilde{i}}]}]$, which can be seen as the *mean* or *template* of class $\tilde{i}$. The *variance of sequence class* $\tilde{i}$ is defined as the sum of all component state variances:

$$\mathbf{\Gamma}^{\tilde{i}} = \sum_{l=1}^{M_{\tilde{i}}} p^{\tilde{i}}_{[l]} \mathbf{\Gamma}^{\tilde{i}}_{[l]} \tag{2}$$

$p^{\tilde{i}}_{[l]}$ is the priori probability of the *l*-th state in sequence model $\mathbf{S}^{\tilde{i}}$, which can be estimated as

$$p^{\tilde{i}}_{[l]} = N^{\tilde{i}}_{[l]} / \sum_{\tilde{i}} N^{\tilde{i}}_{[l]} \tag{3}$$

For concatenate sequences, it is not clear what vectors belong to which class. However, to train an HMM for each class, no pre-segmentation of concatenate sequence into individual sequences is needed. The strategy is to build an HMM for each sequence class, and the concatenate level model is obtained by concatenating the component HMMs. By the so-called cross training, HMMs are able to iteratively refine the segmentation path by aligning vectors of frames to their corresponding models and re-estimate the model parameters. The alignment between vectors and states of HMMs is obtained as a by-product of this process. The following extraction of mean sequences and variances are the same as in the case of individual sequences.

3.2. Objective function and optimization

DR for vector sequences aims at finding a linear transformation $\mathbf{W} \in \mathbb{R}^{\mathbf{d} \times \mathbf{d}'}$, by which the dimension d of original vector $\mathbf{u}^{\tilde{i}}_{n(p)} \in \mathbb{R}^d$ is reduced to d': $\mathbf{y}^{\tilde{i}}_{n(p)} = \mathbf{W}^T \mathbf{u}^{\tilde{i}}_{n(p)} \in \mathbb{R}^{d'}$. so original sequences $\{\mathbf{U}^{\tilde{i}}_n | \mathbf{U}^{\tilde{i}}_n = [\mathbf{u}^{\tilde{i}}_{n(1)}, \mathbf{u}^{\tilde{i}}_{n(2)}, \cdots \mathbf{u}^{\tilde{i}}_{n(P^{\tilde{i}}_n)}] \in \mathbb{R}^{d \times P^{\tilde{i}}_n}, n = 1, \cdots, N_{\tilde{i}}, \tilde{i} = \tilde{1}, \cdots, \tilde{C}\}$ are transformed into $\{\mathbf{Y}^{\tilde{i}}_n | \mathbf{Y}^{\tilde{i}}_n = [\mathbf{y}^{\tilde{i}}_{n(1)}, \mathbf{y}^{\tilde{i}}_{n(2)}, \cdots \mathbf{y}^{\tilde{i}}_{n(P^{\tilde{i}}_k)}] \in \mathbb{R}^{d' \times P^{\tilde{i}}_n}, n = 1, \cdots, N_{\tilde{i}}, \tilde{i} = \tilde{1}, \cdots, \tilde{C}\}$.

We hope that sequences of the same class are projected together while those of different classes are projected as far as possible. In this vein, LSDA determines a linear transformation by maximizing the Fisher criterion such that the separability between sequence classes is maximized at the same time the within-sequence class scatter is minimized.

The sum of pairwise DTW distances between mean sequences of classes can be considered as a measure of separability between sequence classes. We define the transformed mean sequence of class $\tilde{i}$ as: $\mathbf{W}^T \mathbf{m}^{\tilde{i}} = [\mathbf{W}^T \mathbf{m}^{\tilde{i}}_{[1]}, \mathbf{W}^T \mathbf{m}^{\tilde{i}}_{[2]}, \cdots, \mathbf{W}^T \mathbf{m}^{\tilde{i}}_{[L_{\tilde{i}}]}]$. $\mathbf{\Gamma}_w$ is denoted as the within-sequence class scatter, which is defined as *the average variance*:

$$\mathbf{\Gamma}_w = \sum_{\tilde{i}=1}^{\tilde{C}} p^{\tilde{i}} \mathbf{\Gamma}^{\tilde{i}} \tag{4}$$

$p^{\tilde{i}}$ is the priori probability of sequence class $\tilde{i}$, which can be

estimated as

$$p^{\tilde{i}} = N_{\tilde{i}} / \sum_{\tilde{i}} N_{\tilde{i}} \tag{5}$$

Thus this intuition can be formalized as:

$$\begin{aligned} &\max_{\mathbf{W}} \ (tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W}))^{-1}(\sum_{\tilde{i}}\sum_{\tilde{j}} DTW(\mathbf{W}^T\mathbf{m}^{\tilde{i}}, \mathbf{W}^T\mathbf{m}^{\tilde{j}})) \\ &s.t. \ \ \mathbf{W}^T\mathbf{W} = \mathbf{I}_{d'} \end{aligned} \tag{6}$$

Expanding the DTW distance in (6), then (6) can be rewritten as:

$$\begin{aligned} &\max_{\mathbf{W}} \ \left(tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W})\right)^{-1} \\ &\qquad \times \left(\sum_{\tilde{i}}\sum_{\tilde{j}} \min_{\pi^{\tilde{i}},\pi^{\tilde{j}}} \sum_{t=1}^{T} \left\| \mathbf{W}^T\mathbf{m}^{\tilde{i}}_{[\pi_t^{\tilde{i}}]} - \mathbf{W}^T\mathbf{m}^{\tilde{j}}_{[\pi_t^{\tilde{j}}]} \right\|^2\right) \\ \Leftrightarrow &\max_{\mathbf{W}} \ \min_{\pi} \ (tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W}))^{-1} \\ &\qquad \times \left(\sum_{\tilde{i}}\sum_{\tilde{j}} \sum_{t=1}^{T} \left\| \mathbf{W}^T\mathbf{m}^{\tilde{i}}_{[\pi_t^{\tilde{i}}]} - \mathbf{W}^T\mathbf{m}^{\tilde{j}}_{[\pi_t^{\tilde{j}}]} \right\|^2\right) \end{aligned}$$

It is equivalent to the following matrix form:

$$\begin{aligned} &\max_{\mathbf{W}} \ \min_{\mathbf{\Pi}} \ \left(tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W})\right)^{-1} \\ &\qquad \times \left(\sum_{\tilde{i}}\sum_{\tilde{j}} \left\| \mathbf{W}^T\mathrm{M}^{\tilde{i}}\mathbf{\Pi}^{\tilde{i}} - \mathbf{W}^T\mathbf{M}^{\tilde{j}}\mathbf{\Pi}^{\tilde{j}} \right\|^2_{Frob}\right) \\ &= \left(tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W})\right)^{-1} tr\left(\mathbf{W}^T\mathbf{B}_s(\mathbf{\Pi})\mathbf{W}\right) \end{aligned} \tag{7}$$

Where $\|\cdot\|_{Frob}$ is the Frobenious norm. $\mathbf{\Pi}^{\tilde{i}} \in \{0,1\}^{L_{\tilde{i}}\times T}$ and $\mathbf{\Pi}^{\tilde{j}} \in \{0,1\}^{L_{\tilde{j}}\times T}$ are binary alignment indication matrices, whose elements are set as follows: $\mathbf{\Pi}^{\tilde{z}}(\pi_t^{\tilde{z}}, t) = 1$, for $t \in \{1, \cdots, T\}$ and 0 otherwise, for $\tilde{z} = \tilde{i}, \tilde{j}$. The functions of indication matrixes are to replicate the columns of data matrixes associated with the aligned path, resulting in the t-th column of the matrix $\mathbf{W}^T\mathbf{M}^{\tilde{i}}\mathbf{\Pi}^{\tilde{i}} - \mathbf{W}^T\mathbf{M}^{\tilde{j}}\mathbf{\Pi}^{\tilde{j}}$ equals the difference between correspondence vectors at step t. $\mathbf{B}_s(\mathbf{\Pi})$ is defined as: $\mathbf{B}_s(\mathbf{\Pi}) = \sum_{\tilde{i}}\sum_{\tilde{j}} \mathbf{B}(\mathbf{\Pi}^{\tilde{i}}, \mathbf{\Pi}^{\tilde{j}})$ and $\mathbf{B}(\mathbf{\Pi}^{\tilde{i}}, \mathbf{\Pi}^{\tilde{j}}) = (\mathbf{M}^{\tilde{i}}\mathbf{\Pi}^{\tilde{i}} - \mathbf{M}^{\tilde{j}}\mathbf{\Pi}^{\tilde{j}})(\mathbf{M}^{\tilde{i}}\mathbf{\Pi}^{\tilde{i}} - \mathbf{M}^{\tilde{j}}\mathbf{\Pi}^{\tilde{j}})^T$. $\mathbf{B}_s(\mathbf{\Pi})$ depends on the set of all possible alignment indication matrixes of all sequence pairs.

It is difficult to directly optimize the non-convex objective function (7), we optimize the following approximate problem instead:

$$\max_{\mathbf{W}} \ \left(tr(\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W})\right)^{-1} tr\left(\mathbf{W}^T\mathbf{B}_s(\mathbf{\Pi}^*)\mathbf{W}\right) \tag{8}$$

Where $\mathbf{\Pi}^* = \{(\mathbf{\Pi}^{\tilde{i}^*}, \mathbf{\Pi}^{\tilde{j}^*}), 1 \le \tilde{i} < \tilde{j} \le \tilde{C}\}$, and $(\mathbf{\Pi}^{\tilde{i}^*}, \mathbf{\Pi}^{\tilde{j}^*})$ is the optimal alignment indication matrixes found by DTW in original space. This approximation can be seen as a strong pruning, which only keeps the best aligned path. $\mathbf{B}_s(\mathbf{\Pi}^*)$ can be considered as *the between sequence scatter matrix*.

Objective function (8) is a trace ratio problem and no closed-form solution exists. A commonly used solution is to transform such problems to a simpler but inexact ratio trace problem:

$$\max_{\mathbf{W}} \ \ tr((\mathbf{W}^T\mathbf{\Gamma}_w\mathbf{W})^{-1}\mathbf{W}^T\mathbf{B}_s(\mathbf{\Pi}^*)\mathbf{W}) \tag{9}$$

This objective function has the same form as LDA. It can be proved that the columns of $\mathbf{W}$ equal the eigenvectors of Matrix $\mathbf{\Gamma}_w^{-1}\mathbf{B}_s(\mathbf{\Pi}^*)$ corresponding to the d largest eigenvalues, thus the optimization problem (9) boils down to an eigenvalue decomposition problem.

3.3. Relationship with LDA

LDA can be seen as a special case of LSDA by considering non-sequential vectors as sequences of only one vector. In this case, an HMM with only one state is trained for each class, which means that all vectors are aligned to that state. The mean sequence and variance are essentially the mean vector and variance of that class. For any two classes, there is only one possible aligned path between two mean sequences both contain one vector, so the between-sequence scatter is actually the between-class scatter defined in LDA. Thus LSDA degenerates into traditional LDA.

LSDA in turn can be seen as LDA performed on a special model pseudo-space, where each point represents an HMM states sequence. The metric in this pseudo-space is DTW distance instead of Euclid distance and variance is the sum of variances of all component states. It is not really a metric space because DTW distance violates the triangle law.

When an HMM with only one state is trained for each class whose samples are sequences, LSDA degenerates into the method that views all vectors in a sequence as individual samples with the same class label.

Those extensions of LDA can also be modified to improve LSDA in a similar way. For example, the covariance of each state can be attached to the corresponding component of the related mean sequence, and the Euclidean distance can be replaced with the Chernoff distance according to [13] when calculating the DTW distance to tackle heteroscedastic data. Each sequence class can also be divided into subclasses following the idea of [31] to adapt to various data distributions.

3.4. Discussion

- The model topology is not necessarily fixed as in section 3.1. It can be built according to a prior knowledge and the way to process the data. We further note that the mean sequence and variance can be extracted by other generative models, such as cluster generative statistical dynamic time warping [2].

- The between sequence scatter matrix $\mathbf{B}_s(\mathbf{\Pi}^*)$ can be decomposed as follows:

$$\mathbf{B}_s(\mathbf{\Pi}^*) = \sum_{\tilde{i}} \sum_{\tilde{j}} \mathbf{B}_{\tilde{i},\tilde{j}} = \sum_{\tilde{i}} \sum_{\tilde{j}} \sum_{t=1}^{T} \mathbf{B}_{\tilde{i},\tilde{j},t}$$

$$\mathbf{B}_{\tilde{i},\tilde{j},t} = \left(\mathbf{m}^{\tilde{i}}_{[\pi^{\tilde{i}}_t]} - \mathbf{m}^{\tilde{j}}_{[\pi^{\tilde{j}}_t]}\right)\left(\mathbf{m}^{\tilde{i}}_{[\pi^{\tilde{i}}_t]} - \mathbf{m}^{\tilde{j}}_{[\pi^{\tilde{j}}_t]}\right)^T$$

  Each $\mathbf{B}_{\tilde{i},\tilde{j},t}$ is spanned by the two mean vectors of the state pair aligned in the t-th step of DTW. We can see that only those state pairs that can reflect the difference between two sequence models are used to calculate the matrix $\mathbf{B}_s(\mathbf{\Pi}^*)$. The difference of two states from the same model will not contribute to $\mathbf{B}_s(\mathbf{\Pi}^*)$.

- The definition of within-class scatter (4) is equivalent to that of LDA by considering each state as a separate class. This can be seen by substituting (2), (3) and (5) to (4) as shown in the following, which is the average of all within-state variances from all models:

$$\mathbf{\Gamma}_w = \sum_{\tilde{k}=1}^{\tilde{C}} p^{\tilde{k}} \mathbf{\Gamma}^{\tilde{k}} = \sum_{\tilde{k}=1}^{\tilde{C}} \sum_{l=1}^{M_{\tilde{i}}} (N^{\tilde{i}}_{[l]} / \sum_{\tilde{k}} N_{\tilde{k}}) \mathbf{\Gamma}^{\tilde{i}}_{[l]}$$

  $N^{\tilde{i}}_{[l]} / \sum_{\tilde{k}} N_{\tilde{k}}$ is the prior of the l-th state of class $\tilde{i}$.

- Although HMMs are first trained to obtain statistics, LSDA is just a dimensionality reduction method and is irrelevant to the subsequent classifier which is not limited to HMM. As long as the transformation is obtained, any method can be used to do classification. For instance, two classifiers, HMM and DTW, are used as classifiers for individual sequences in section 4. For concatenate sequences, since both segmentation and recognition are needed, only HMM is adopted.

- The process of HMM training and statistics extraction is necessary to perform whether LDA or LSDA for vector sequences. DTW only need to be implemented once for each sequence class pair with respect to LSDA, the computation cost increases linearly with the number of states for each HMM. However, since each state is considered as a class with LDA, additional computation is brought in quadraticly when calculating the between-state scatter. Thus the computation cost of LSDA is at least comparable to that of LDA.

4. Experiments on individual sequences

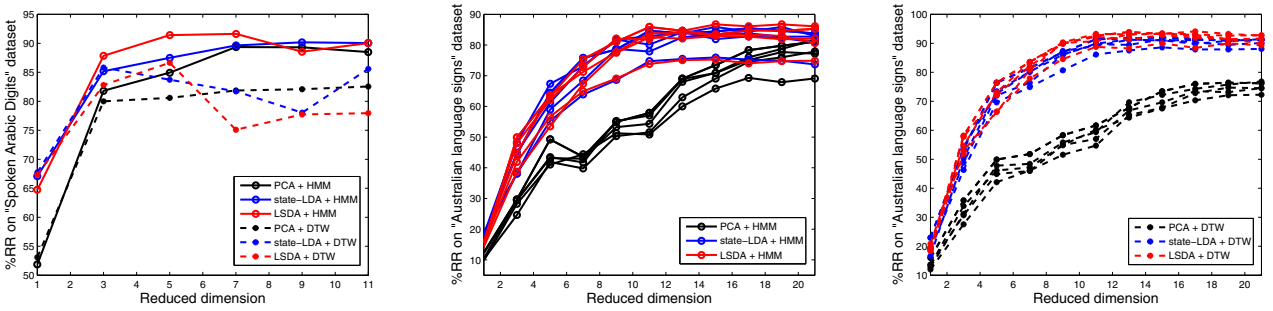
In this section we evaluate the effectiveness of LSDA on individual sequences. Experiments are carried out on two individual sequence datasets from the UCI Machine Learning Repository [17].

Dataset. The Spoken Arabic Digits dataset consists of 8800 vector sequences from ten classes. The vectors in sequences contain 13 mel-frequency cepstrum coefficients. 44 mails and 44 females native Arabic speakers repeated the ten digits (from 0 to 9) ten times. The length of sequences varies from 4 to 93 frames. The High-quality recordings of Australian sign language signs (HAS) dataset [12] contains 2565 vector sequences of Auslan (Australian Sian Language) signs captured from a native signer using high-quality position trackers. There are totally 95 different signs with 27 samples per sign. The length of sample sequences varies with an average length of approximately 57 frames, and 22 attributes were extracted from each frame.

Experimental setup. For the Spoken Arabic Digits dataset, the data has already been divided into a training set of 6600 samples with 660 samples per class and a test set of 2200 samples with 220 samples per class. Experiments were carried out following this partition. For the HAS dataset, we divided it into five subsets, of which four were used for training and the remaining one for test. Experiments were carried out on such fivefold validation. For both datasets, we compared the proposed LSDA to two baselines: unsupervised PCA and supervised LDA by considering each state as individual classes as introduced in section 2 (denoted as state-LDA).

Classification in the transformed subspace was performed by two classifiers: DTW and HMM. For DTW, pairwise DTW distances are calculated within the same class, and the sample which has the minimum distance with all other samples is selected as the template of the class. A new test sequence is matched to templates of all classes, and the class of the template with the highest similarity is determined as the label of the sequence. For HMM, an HMM is built for each sequence class. The likelihood of a new sequence is estimated by each HMM, and the class related to the HMM with the highest score is assigned as the label. HTK [28] was used to perform training and decoding.

Results. Fig. 2(a) shows the recognition rate (RR) of the three methods by using the two classifiers. Fig. 2(b) and 2(c) show the RR of 5 splits obtained by the HMM and DTW classifiers as a function of the dimensionality, respectively. The dimension of vectors was reduced to all possible odd numbers. A primary HMM with 4 states and a mixture of 5 Gaussian densities for each state was built to obtain statistics. HMMs for classification in the subspace had the same topology with primary HMMs. It can be observed on the three figures that supervised LSDA and state-LDA outperform unsupervised PCA on all these dimensions. The variances of performances by LSDA and state-LDA are comparable, but the proposed LSDA generally outperforms state-LDA on most dimensions. For both classifiers, the best results obtained by LSDA among these dimensions are also better than those by state-LDA on both datasets.



(a) (b) (c)

Figure 2. Recognition rate on individual sequence datasets. (a) Performance by HMM and DTW on the "Spoken Arabic Digits" dataset; (b) performance of 5 splits by HMM on the HAS dataset; (c) performance of 5 splits by DTW on the HAS dataset.

5. Experiments on concatenate sequences

5.1. APTI Arabic printed text database

Dataset. APTI [23] is a synthetic Arabic printed text image database which contains more than 45 million Arabic words from a lexicon of 113,284 different words. The images are synthetically created in low resolution. In our experiments, we use parts of APTI database with font "Arabic Transparent", style "Plain" and size "24" . 5 sets are available for these parts, 75,750 images belonging to set 1 to 4 were used for training and set 5 of additional 18,868 images was used for test.

Feature extraction. Each word image was rescaled to 30 rows in height, while the aspect ratio was maintained. A window was slid on the image from right to left following the Arabic writing direction and a feature vector was extracted within each window. Thus the image was represented by a vector sequence. The width of the window was set to be 10, and the movement was set to be 1. In each window, the vertical repositioning technology [1] was used to deal with vertical distortion and a set of 79 features was extracted using the method proposed in [24]. The feature set was a combination of log-space distribution features and several other baseline-independent features.

Training and recognition. We built an HMM for every character. Each character model had a right to left topology with self-loops and one state skipping transitions permitted, which was shared by HMMs for both alignment and classification. There were 120 models in total to model different shapes of characters as well as additional marks. The character model grouping method [22] was adopted to merge similar visual glyphs. 65 character models remained after grouping. The dimension of features was reduced to from 10 to 50 with interval of 5, respectively. Then the HMM-based training and lexicon-free recognition were performed with 4 states for each character model and a mixture of 3 Gaussians per state. No language model was used.

Results. Table 1 shows the performance of the PCA, state-LDA and LSDA. Two performance measures were calculated: the word recognition rate and the character accuracy rate. The word recognition rate is the percentile of completely recognized words, which means all component characters and their order are correctly recognized. The character accuracy rate is calculated as follows: the total number of labels in reference transcriptions minus the substitution errors, deletion errors and insertion errors, and then divided by the total number of character labels. The best performances of the three measures with corresponding dimensions are shown in the Best column of these tables. From these results, we can see that in most reduced dimensions, LSDA outperforms state-LDA and PCA. On these two measures, the best results achieved by LSDA are better than those by state-LDA and PCA.

5.2. IFN/ENIT Arabic handwriting database

Dataset. The public part of the IFN/ENIT benchmark database (v2.0ple) [18] is divided into five sets labeled from *a* to *e*, with the numbers of binary word images in each set are 6537, 6710, 6477, 6735 and 6033, respectively. The total 32492 words from a lexicon of 937 Tunisian town/village names are written by more than 1000 writers.

Feature extraction. The same feature extraction method as in the APTI database was applied here with different parameters. Each word image was rescaled to 80 rows in height.The width of the window was set to be 30, and the movement each time was set to be 2. A set of 103 features was extracted in each window which is different from that in APTI database since the dimensionality of features is window-size dependent.

Training and recognition. The same HMM-based training and recognition strategy as in APTI database was applied again with different topologies. The state number adaptation technique [1] was adopted to deal with the significantly different average lengths and structural complexities of different characters. The number of states after adaptation was set to be the average segment length multiplied

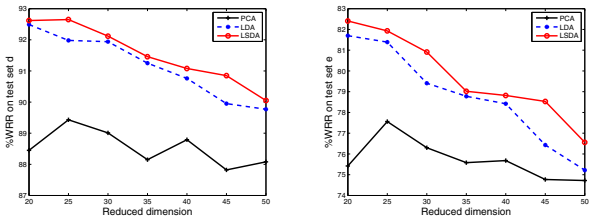
Dim(d)	10	15	20	25	30	35	40	45	50	Best
PCA	69.53	76.05	77.99	77.09	77.81	78.04	76.43	74.37	75.13	78.04 (35)
state-LDA	77.64	82.56	82.69	82.55	82.25	81.66	82.62	83.91	82.97	83.91 (45)
LSDA	80.66	79.98	83.02	82.04	81.06	84.30	83.08	**85.05**	80.20	**85.05** (45)

(a) % Word correct rate

Dim(d)	10	15	20	25	30	35	40	45	50	Best
PCA	93.20	94.98	95.32	95.20	95.32	95.17	94.82	94.50	94.55	95.32 (20,30)
state-LDA	95.37	96.55	96.64	96.65	96.54	96.46	96.62	96.78	96.61	96.78 (45)
LSDA	96.07	96.03	96.68	96.48	96.29	96.85	96.61	**96.95**	96.09	**96.95** (45)

(b) % Character accuracy rate

Table 1. (a) Word recognition rate and (b) character accuracy rate on APTI database with font "Arabic Transparent", style "Plain" and size "24".



(a) (b)

Figure 3. Word recognition rate on (a) test set *d* and (b) test set *e* of the IFN/ENIT database.

Systems	%WRR training set-test set		
	abc-d	abc-e	abcd-e
RWTH OCR [6]	96.53	—-[a]	92.74
UPV PRHLT [1]	95.3	—-	93.9
LSDA+HMM	**97.18**	93.47	**93.97**

Table 2. Comparison of several advanced Arabic handwriting recognition systems on the IFN/ENIT dataset. [a]'—' denotes no results reported on that partition.

by an alignment factor. Preliminary HMMs for obtaining statistics and final HMMs for lexicon-driven recognition shared the same topology with various numbers of states and 5 Gaussian densities per state.

Performance Comparison with Different dimensionality reduction Methods. A sub-set of 66 features from the total 103 features were used to reduce the scale of the problem and accelerate the recognition speed by removing all concavity, structure and point features and following a rough zoning division, since the experimental setup aims to evaluate the relative performances of different dimensionality reduction methods. The total 178 character models were grouped into 98 models, and the number of states after adaptation was set to 0.2 times the average segment length.

Fig. 3 shows the word recognition rate on test set *d* and test set *e* by using set *abc* of the IFN/ENIT database as training sets. It is observed again that on most dimensions LSDA archieves better results and the best results are obtained by LSDA on both test sets.

Results Compared with Other Arabic Handwritten Recognition Systems. In this subsection we compare the results with some advanced systems. All the 103-dimensional features were used and no character models were grouped. The state labels of vectors were generated by preliminary HMMs trained with the original features, and then the dimension of features were reduced to 30. The factor in the state number adaptation technique was set to be 0.7. The final HMMs were trained with various numbers of states and 5 Gaussian densities per state to perform recognition on the commonly used partitions abc-de and abcd-e of the IFN/ENIT dataset.

The word recognition rates of our method together with two Arabic handwriting systems are shown in table 2. These systems for comparison include RWTH-OCR and UPV PRHLT - the winner of ICDAR 2011 and ICFHR 2010 Arabic handwriting competition [15, 14], respectively. In these competitions set *a* to *e* from the IFN/ENIT database were used as training sets and set *f* and *s* were used as test sets. The two test sets *f* and *s* are not public so research institutions (including those who participate competitions) can only use set *a* to *e* for evaluation. Results reported in [15, 14] on set *d* and *e* were obtained by training on set *a* to *e*, which represented performance on training set. The results we compared were from [6] and [1] which were reported by these groups themselves on partitions $abc-d$ and $abcd-e$ after ICDAR 2011 competition. All these systems adopted dimensionality reduction methods. The proposed LSDA followed by a relatively simple HMM-based classier achieves comparable results with these systems.

6. Conclusions

In this paper we have presented a model-based dimensionality reduction method for vector sequences, LSDA,

which projects vectors in sequences into a subspace such that sequences as a whole can be maximally separated into different classes. LSDA can discover differences and improve discrimination in the latent space between sequence classes. LDA can be considered as a special case of LSDA, while LSDA can be seen as LDA performed in a pseudo-model space where points represent sequences by using DTW distance instead of Euclidean distance. We have demonstrated the effectiveness of LSDA in classifying both individual sequence data and concatenate sequence data. LSDA outperforms unsupervised PCA and supervised state-LDA in both cases on several different datasets. In our future work, we intend to extend LSDA to nonlinear cases by kernelization.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 60933010 and 61032008.

References

[1] I. Alkhoury, A. Gimenez, and A. Juan. *Guide to OCR for Arabic scripts*, chapter Arabic handwriting recognition using bernoulli HMMs. Springer, 2012.

[2] C. Bahlmann and H. Burkhardt. The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping. *PAMI*, 26(3):299–310, 2004.

[3] A. Bharath and S. Madhvanath. Hmm-based lexicon-driven and lexicon-free word recognition for online handwritten indic scripts. *PAMI*, 34(4):670–682, 2012.

[4] H. Cao, R. Prasad, and P. Natarajan. Ocr-driven writer identification and adaptation in an hmm handwriting recognition system. In *ICDAR*, 2011.

[5] K. Chan and A. Fu. Efficient time series matching by wavelets. In *ICDE*, 1999.

[6] P. Dreuw, D. Rybach, G. Heigold, and H. Ney. *Guide to OCR for Arabic scripts*, chapter RWTH OCR: A large vocabulary optical character recognition system for Arabic scripts. Springer, 2012.

[7] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *SIGMOD*, 1994.

[8] J. Frank, S. Mannor, J. Pineau, and D. Precup. Time series analysis using geometric template matching. *PAMI*, 35(3):740–754, 2013.

[9] A. Geiger, R. Urtasun, and T. Darrell. Rank priors for continuous non-linear dimensionality reduction. In *CVPR*, 2009.

[10] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *ICASSP*, 1992.

[11] O. Hamsici and A. Martinez. Bayes optimality in linear discriminant analysis. *PAMI*, 30(4):647–657, 2008.

[12] M. W. Kadous. *Temporal classification: extending the classification paradigm to multivariate time series*. PhD Thesis (draft), School of Computer Science and Engineering, University of New South Wales, 2002.

[13] M. Loog and R. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *PAMI*, 26(6):732–739, 2004.

[14] V. Märgner and H. Abed. Icfhr 2010 arabic handwriting recognition competition. In *ICFHR*, 2010.

[15] V. Märgner and H. Abed. Icdar 2011 arabic handwriting recognition competition. In *ICDAR*, 2011.

[16] R. A.-H. Mohamad, L. Likforman-Sulem, and C. Mokbel. Combining slanted-frame classifiers for improved hmm-based arabic handwriting recognition. *PAMI*, 31(7):1165–1177, 2009.

[17] D. Newman, S. Hettich, C. Blake, and C. Merz. *UCI repository of machine learning databases*. Dept. of Information and Computer Sciences, Univ. of California, Irvine, http://www.ics.uci.edu/ mlearn/MLRepository.html, 1998.

[18] M. Pechwitz, S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri. Ifn/enit-database of handwritten arabic words. In *CIFED*, 2002.

[19] J. Rodríguez-Serrano and F. Perronnin. A model-based sequence similarity with application to handwritten word spotting. *PAMI*, 34(11):2108–2120, 2012.

[20] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *TASSP*, 26(1):43–49, 1978.

[21] A. Shyr, R. Urtasun, and M. Jordan. Sufficient dimension reduction for visual sequence classification. In *CVPR*, 2010.

[22] F. Slimane, R. Ingold, S. Kanoun, A. Alimi, and J. Hennebert. Impact of character models choice on arabic text recognition performance. In *ICFHR*, 2010.

[23] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert. A new arabic printed text image database and evaluation protocols. In *ICDAR*, 2009.

[24] B. Su, X. Ding, L. Peng, and C. Liu. A novel baseline-independent feature set for arabic handwriting recognition. In *ICDAR*, 2013.

[25] Y. Ukrainitz and M. Irani. Aligning sequences and actions by maximizing space-time correlations. In *ECCV*, 2006.

[26] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 30(2):283–298, 2008.

[27] D. You, O. Hamsici, and A. Martinez. Kernel optimization in discriminant analysis. *PAMI*, 33(3):631–638, 2011.

[28] S. Young, G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*. Cambridge University Engineering Dept., 2001.

[29] F. Zhou and F. D. la Torre. Canonical timewarping for alignment of human behavior. In *NIPS*, 2009.

[30] F. Zhou and F. D. la Torre. Generalized time warping for multi-modal alignment of human motion. In *CVPR*, 2012.

[31] M. Zhu and A. Martinez. Subclass discriminant analysis. *PAMI*, 28(8):1274–1286, 2006.