Learning Low-Dimensional Temporal Representations
Supplementary Material

Bing Su\textsuperscript{1} Ying Wu\textsuperscript{2}

1. Proof of Lemma 1

Lemma 1. Objective function (#3)\textsuperscript{1} is equivalent to the trace maximization problem

\[
\max \text{tr}(F^T Z^T Z F).
\]

Proof. It turns out that the \(k\)-th column of \(ZFF\) is exactly \(m_k\) (i.e., the mean of the \(j\)-th stage of class \(c\)) that the \(k\)-th column \(z_k\) of \(Z\) is aligned to. Therefore, the objective function (#3) can also be written as

\[
\min \|Z - ZFF\|_F^2 \quad \Leftrightarrow \min \text{tr}(Z^T Z) - \text{tr}(F^T Z^T Z F) \quad \Leftrightarrow \max \text{tr}(F^T Z^T Z F)
\]

2. Proof of Lemma 2

Lemma 2. Objective function (#7) is equivalent to the trace maximization problem

\[
\max \text{tr}(F^T Z^T W (W^T Z Z^T W)^{-1} W^T Z F).
\]

Proof. If the data has zero-centered, \(\sum_i z_i/N_t = 0\), (otherwise we can remove the overall mean of all the vectors in all the training sequences), then \(S_t\) and \(S_b\) can be reformulated as

\[
S_t = ZZ^T,
\]

\[
S_b = \sum_{i=1}^C \sum_{u=1}^L p_{iu} m_i m_i^T = ZFF^T Z.
\]

Since \(S_b + S_w = S_t\), the objective (#7) is equivalent to

\[
\max \text{tr}((W^T S_w W)^{-1} W^T S_b W) \Leftrightarrow \max \text{tr}((W^T S_t W)^{-1} W^T S_b W),
\]

which can further be reformulated as

\[
\text{tr}((W^T S_t W)^{-1} W^T S_b W) = \text{tr}((W^T Z Z^T W)^{-1} W^T ZFF^T Z W) = \text{tr}(F^T Z^T W (W^T Z Z^T W)^{-1} W^T Z F)
\]

3. Proof of Theorem 1

Theorem 1. The LT-LDA algorithm (Alg. 2) is guaranteed to converge.

Proof. In the first stage, LT-LDA optimizes over \(P\) by fixing \(W\) using Alg. 1. According to Lemma 1, this actually learns the abstract template from the whitened and projected feature sequences by optimizing \(\max \text{tr}(F^T Z^T Z F)\). Let \(\hat{Z} = \Gamma_w^{-1} W^T Z\) be the whitened and projected data matrix, \(\Gamma_w = W^T Z Z^T W\) is the total scatter of the projected data. Thus we have

\[
\text{tr}(F^T \hat{Z}^T Z F) = \text{tr}(F^T Z^T W (W^T Z Z^T W)^{-1} W^T Z F)
\]

which is exactly the objective Eq. (2).

In the second stage, LT-LDA optimizes over \(W\) for given \(P\). According to Lemma 2, this also optimizes Eq. (2). Both iterative stages decrease the objective value of Eq. (2) monotonically. Since \(F\) is an orthogonal matrix, the objective (2) is bounded from above. This guarantees the convergence of Alg. 2.

4. Proof of Theorem 2

Theorem 2. Let \(G = Z^T Z\) be the Gram matrix. When the dimensionality is reduced to a specific value \(d^* = \frac{2}{\sqrt{d'}}\)
min(CL, d, N_t) and a regularization term δI_N_t is added to the total scatter S_t, where I_N_t is the N_t-order identity matrix, if W* and F* are the optimal solutions of the trace maximization problem (2):

$$\max_{W,F} tr(F^T Z^T W (W^T (Z Z^T + δI_{N_t}) W)^{-1} W^T Z F)$$

(4)

then F* is also the optimal solution of the problem

$$\max_{F} tr(F^T (I_{N_t} - (I_{N_t} + \frac{1}{δ} G)^{-1}) F)$$

(5)

Proof. We follow the proof in (Ye et al., 2007). According to the representer theorem (Schölkopf & Smola, 2002), the optimal projection matrix W has the form W = ZA, where A is a coefficient matrix. The objective (4) is transformed to

$$tr(F^T Z^T Z A (A^T Z^T (Z Z^T + δI_{N_t}) Z A)^{-1} A^T Z^T Z F)$$

$$= tr(F^T G A (A^T (G G + δG) A)^{-1} A^T G F)$$

$$= tr(A^T G F)$$

(6)

By defining Γ_b = G F^T G and Γ_w = G G + δG, we can find that Eq. (6) has a similar form with the generalized LDA problem (Ye, 2005), which can be solved by constructing a matrix Q that simultaneously diagonalizes Γ_b and Γ_w. Q is constructed as follows.

G is symmetric and positive semi-definite. If the d-dimensional features are not linearly dependent (otherwise we can remove the linearly correlated dimensions), the rank is of G is r = min(d, N_t). The SVD of G has the form

$$G = U \Sigma U^T = U_r \Sigma_r U_r^T,$$

where U is an orthogonal and square matrix, Σ = diag(λ_1,⋯,λ_r,0,⋯,0), U_r consist of the first r columns of U, and Σ_r = diag(λ_1,⋯,λ_r) contains only the non-zero singular values.

Define V = (Σ_r^2 + δΣ_r)^{-1/2} Σ_r U_r^T F. The SVD of V is denoted as V = MV N^T, where M and N are orthogonal matrices, Σ_V is a diagonal matrix of d'-th order, and d' is the rank of V. d' = min(rank(S_b), r) = min(CL, d, N_t).

By constructing Q as

$$Q = U diag((Σ_r^2 + δΣ_r)^{-1/2} M, I_{N_t-r}),$$

Γ_b and Γ_w are simultaneously diagonalized by Q.

$$Q^T Γ_b Q = diag(Σ_r^2, 0_{N_t-r}),$$

$$Q^T Γ_w Q = diag(I_r, 0_{N_t-r}).$$

The equality holds when A = A*, where the dimensionality is implicitly reduced to d'.

5. Experimental Setup on Another Dataset

“Spoken Arabic Digits (SAD)” dataset from the UCI Machine Learning Repository (Bache & Lichman, 2013). This dataset consists of 8,800 vector sequences from ten classes. The vectors in sequences are 13 mel-frequency cepstrum coefficients (MFCCs) corresponding to spoken Arabic digits. The ten spoken Arabic digits were repeated ten times by 44 males and 44 females Arabic native speakers and hence there are 880 sequence samples per digit class. The length of sequences varies from 4 to 93 frames. The dataset has already been split into training and test sets, where 660 samples from each class are used for training and the remaining 220 sequences are used for testing.

For the SAD dataset, each spoken digit has also been represented by a sequence of 13-dimensional frame-wise MFCC features.

6. Influence of Parameters

In this section we evaluate the influence of the parameters on the ChaLearn gesture dataset. Similar to the MSR Action3D dataset, different performance measures including accuracy, MAP, and F-score with the rank pooling and the SVM classifier are evaluated by increasing L from 3 to 21.
with an interval of 3 while fixing \( a \) to 2, and increasing \( a \) from 1 to 5 with an interval of 0.5 while fixing \( L \) to 8, respectively. The results on the dataset are shown in Fig. 1, where the reduced dimensionality is fixed to 45.

The optimal parameters are generally the same for multi-class indicators including accuracy, precision, recall and F-score, but are different for MAP. On the ChaLearn dataset, LT-LDA is insensitive to \( L \) for multi-class indicators. When \( L = 12 \), LT-LDA achieves the highest MAP. ATs with more stages help to discriminate the subtle differences among fine gesture actions in this dataset. LT-LDA is also not very sensitive to \( a \) on this dataset. It seems that allowing larger warping leads to better results because the subtle differences can be captured more easily by more flexible alignments.

7. Effects of the Joint Learning of the Subspace and the Latent Alignments

In this section we compare LT-LDA and ini-LT-LDA on more datasets. On both the ChaLearn dataset and the MSR Action3D dataset, the frame-wide features are based on the relative 3D joint positions, and we only show the comparisons on the dataset.

Figure 2. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies with the DTW classifier and (b) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the ChaLearn Gesture dataset.

Figure 4. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies and (b) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the Olympic Sports dataset.

The comparisons of the two algorithms by the SVM classifier on the large scale Olympic Sports dataset are shown in Fig. 4. We can find LT-LDA significantly outperforms ini-LT-LDA by a much larger margin than the Kinect based datasets. This is because the depth information and the locations of human joints are available for the ChaLearn dataset. Thus the alignments in the original space are accurate and should be preserved in the optimal subspace. After the refinement of LT-LDA, the alignments remain nearly unchanged and the performance changes are small. On the Olympic dataset, only the raw videos with complex backgrounds are available. The initial alignments are quite noisy, and refining them by LT-LDA improves the performances significantly.

In summary, in nearly all cases, the best results among all these dimensions of LT-LDA outperform those of ini-LT-LDA. Jointly learning the latent alignments associated with the subspace does help to improve the classification performance in the subspace. This is because the temporal structures and the alignments may change from those in the original space. In the learned subspace of ini-LT-LDA, although different classes get better separated under the alignments in the original space, additional confusions may be introduced due to the change of alignments. While for LT-LDA, since the separability is maximized in the subspace under the corresponding alignments, the learned subspace gets joint optimality among all possible subspaces.

We can also observe that on some datasets with some classifiers, the performances of LT-LDA are lower than those of ini-LT-LDA in some dimensions. This is because the aligned paths of each class are learned without discriminating other classes, and the objective of LT-LDA is not directly related to a classification performance measure. For some dimensions, although our joint learning leads to bet-
Figure 3. Comparisons of the proposed LT-LDA without and with the joint learning of the latent alignments. (a) Accuracies with the HMM classifier (b) Accuracies with the DTW classifier (c) MAPs with the rank pooling and the SVM classifier and (b) Multi-class average recalls with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the SAD dataset.

Figure 5. (a) Accuracies with the HMM classifier (b) accuracies with the DTW classifier and (c) MAPs with the rank pooling and the SVM classifier as functions of the dimensionality of the subspace on the SAD dataset.

8. Comparison with Different Dimensionality Reduction Methods

In this section we compare LT-LDA with other dimensionality reduction methods on the two additional datasets. Fig. 5 depicts the performances by the three classifiers as functions of the dimensionality of the learned subspace on the SAD dataset, respectively. We can observe that the proposed LT-LDA achieves the best performances among all these dimensionality reduction methods by all the three classifiers with different evaluation measures on nearly all the datasets. On the SAD dataset, by the HMM classifier and the SVM classifier, LSDA performs comparatively with LT-LDA. This is because sufficient training samples are available on this dataset, and the dimensionality of frame-wise features is low. Therefore, LSDA can also reliably train HMMs to obtain sequence statistics.

References


